

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

THOMAZ DINIZ PINTO DE MORAIS

INCORPORANDO REGIONALISMOS E SOTAQUES EM MODELOS ABERTOS DE SÍNTESE DE FALA

CAMPINA GRANDE - PB
2026

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Incorporando Regionalismos e Sotaques em Modelos Abertos de Síntese de Fala

Thomaz Diniz Pinto de Morais

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Processamento de Fala e Inteligência Artificial

Prof. Dr. Herman Martins Gomes
(Orientador)

Campina Grande, Paraíba, Brasil
©Thomaz Diniz Pinto de Morais, Agosto de 2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM CIENCIA DA COMPUTACAO
Rua Aprígio Veloso, 882, Edifício Telmo Silva de Araújo, Bloco CG1, - Bairro Universitário, Campina Grande/PB, CEP 58429-900
Telefone: 2101-1122 - (83) 2101-1123 - (83) 2101-1124
Site: http://computacao.ufcg.edu.br - E-mail: secpg@computacao.ufcg.edu.br

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

THOMAZ DINIZ PINTO DE MORAIS

INCORPORANDO REGIONALISMOS E SOTAQUES EM MODELOS ABERTOS DE SÍNTESE DE FALA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 24/08/2026

Prof. Dr. **Herman Martins Gomes**, UFCG, Orientador

Profa. Dra. **Joseana Macêdo Fechine Régis de Araújo**, UFCG, Examinador Interno

Prof. Dr. **Frederico Santos de Oliveira,** UFMT, Examinador Externo

Documento assinado eletronicamente por **HERMAN MARTINS GOMES**, **PROFESSOR 3 GRAU**, em 25/08/2026, às 11:06, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **JOSEANA MACEDO FECHINE**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 11/09/2026, às 15:52, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **6809488** e o código CRC **DD267084**.

Referência: Processo nº 23096.058536/2026-45 SEI nº 6809488

Universidade Federal de Campina Grande - UFCG
Sistema de Bibliotecas - SISTEMOTECA
Catalogação de Publicação na Fonte. UFCG - Biblioteca Central

M827i Morais, Thomaz Diniz Pinto de.

Incorporando regionalismos e sotaques em modelos abertos de síntese de fala / Thomaz Diniz Pinto de Morais. – 2026.

198 f. : il. color.

Dissertação (mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2026.

"Orientação: Prof. Dr. Herman Martins Gomes".

Referências.

1. Processamento da Fala. 2. Inteligência Artificial. 3. Síntese de Fala. 4. Sotaque Regional. 5. Ajuste Fino. 6. Português Brasileiro. 7. Aprendizagem Profunda. I. Gomes, Herman Martins. II. Título.

UFCG/BC CDU 004(043.3)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA ITAPUANA SOARES DIAS GONÇALVES CRB-15/93

Resumo

Modelos abertos de síntese de fala apresentam limitações quanto à diversidade de sotaques regionais do português brasileiro. Esta dissertação investiga a incorporação do sotaque paraibano em três arquiteturas (*XTTS*, *Fish Speech* e *Orpheus TTS*), comparando três estratégias de adaptação: clonagem de voz, ajuste fino e treinamento a partir de pesos aleatórios. Experimentos exploratórios evidenciaram limitações práticas da clonagem de voz e do treinamento a partir de pesos aleatórios, e o ajuste fino, identificado como a alternativa mais promissora, foi validado sistematicamente. Para isso, foi construído o corpus *FALA_PB*, com cerca de 111 horas de fala paraibana organizadas em segmentos curtos, cada um pareado à respectiva transcrição em nível de sentença. O corpus agrega dados de bases pré-existentes (*CoLingPB*, *Braccent* e *Sotaque Brasileiro*) e de uma coleta colaborativa própria; em especial, as entrevistas do *CoLingPB* foram adaptadas para a tarefa de síntese, sendo os áudios segmentados em sentenças, recortados automaticamente com *FFmpeg* e filtrados por qualidade acústica. Os modelos ajustados com esta base foram avaliados por indicadores automáticos de qualidade e inteligibilidade e por avaliação subjetiva com 56 ouvintes. Os resultados indicam que o ajuste fino foi eficaz: o sotaque paraibano foi percebido como mais presente nos modelos ajustados em cerca de 88% a 90% das respostas, sem redução perceptível de qualidade, com destaque para o *XTTS*, que obteve o melhor desempenho geral. Como contribuições da dissertação, são fornecidos todos os artefatos de código e dados desenvolvidos, além de um protocolo reprodutível de avaliação, ampliando a representatividade regional na síntese de fala aberta para o português brasileiro e contribuindo com futuras pesquisas que objetivem a síntese automática de regionalismos e sotaques de outras regiões brasileiras.

Palavras-chave: síntese de fala; sotaque regional; ajuste fino; português brasileiro; aprendizagem profunda.

Abstract

Open-weight speech synthesis models show limitations regarding the diversity of regional accents of Brazilian Portuguese. This dissertation investigates the incorporation of the Paraíba accent into three architectures (*XTTS*, *Fish Speech*, and *Orpheus TTS*), comparing three adaptation strategies: voice cloning, fine-tuning, and training from scratch. Exploratory experiments revealed practical limitations of voice cloning and training from scratch, and fine-tuning, identified as the most promising alternative, was systematically validated. To this end, the *FALA_PB* corpus was built, with about 111 hours of Paraíba-accented speech organized into short segments, each paired with its corresponding sentence-level transcription. The corpus aggregates data from pre-existing databases (*CoLingPB*, *Braccent*, and *Sotaque Brasileiro*) and from a dedicated collaborative collection; in particular, the *CoLingPB* interviews were adapted for the synthesis task, with the audios segmented into sentences, automatically cut with *FFmpeg*, and filtered by acoustic quality. The fine-tuned models were evaluated through automatic indicators of quality and intelligibility and through a subjective evaluation with 56 listeners. The results indicate that fine-tuning was effective: the Paraíba accent was perceived as more present in the fine-tuned models in about 88% to 90% of the responses, with no perceptible loss of quality, with *XTTS* achieving the best overall performance. As contributions of this dissertation, all developed code and data artifacts are provided, in addition to a reproducible evaluation protocol, broadening regional representation in open speech synthesis for Brazilian Portuguese and supporting future research aimed at the automatic synthesis of regionalisms and accents from other Brazilian regions.

Keywords: speech synthesis; regional accent; fine-tuning; Brazilian Portuguese; deep learning.

Agradecimentos

Dedico este trabalho àquela que, desde pequeno, sempre me chamava de mestrinho. Descanse em paz, mãe.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Sumário

1 Introdução **1**

1.1 Contextualização 1

1.2 Motivações 2

1.3 Problema de Pesquisa 3

1.4 Questões de Pesquisa 4

1.5 Hipóteses 4

1.6 Objetivos 5

1.6.1 Objetivo Geral 5

1.6.2 Objetivos Específicos 5

1.7 Contribuições 6

1.8 Declaração de Uso de Inteligência Artificial Generativa 7

1.9 Organização da Dissertação 8

2 Fundamentação Teórica **9**

2.1 Síntese de Fala 9

2.1.1 Evolução Histórica da Síntese de Fala 10

2.1.2 Síntese Concatenativa 11

2.1.3 Síntese Paramétrica 12

2.1.4 Síntese por Aprendizagem Profunda 14

2.2 Redes Neurais Artificiais 16

2.2.1 Redes Densas e o Conceito de Pesos 16

2.2.2 *Transformers* e o Mecanismo de Atenção 17

2.3 Representações de Áudio 17

2.3.1 Forma de Onda, Espectrograma e Escala Mel 17

2.3.2 Representações Latentes e Tokens Acústicos 19
2.4 Componentes de Sistemas de Síntese de Fala com Aprendizagem Profunda 19
2.4.1 Checkpoints e Formato de Armazenamento 20
2.5 Sotaques e Variação Linguística 20
2.5.1 Prosódia e Entonação 21
2.5.2 Variação Fonética 22
2.5.3 Regionalismos no Português Brasileiro 22
2.6 Ferramentas e Modelos Abertos de TTS 23
2.6.1 Pré-Treinamento 23
2.6.2 Ajuste Fino 24
2.6.3 Clonagem de Voz 25
2.7 Avaliação de Sistemas de Síntese de Fala 26
2.7.1 Avaliação Perceptual 26
2.7.2 *Mean Opinion Score* (MOS) 27
2.7.3 Preditores Automáticos de MOS 28
2.7.4 Avaliação Baseada em STT e Distância de Levenshtein 28
2.7.5 Testes de Significância Estatística 30
2.8 Síntese do Capítulo 32

3 Revisão da Literatura **33**
3.1 Metodologia da Revisão Bibliográfica 34
3.1.1 Questões Norteadoras 34
3.1.2 Estratégia de Busca 34
3.1.3 Critérios de Inclusão 34
3.1.4 Critérios de Exclusão 35
3.1.5 Processo de Seleção dos Estudos 35
3.2 Modelos de Síntese de Fala Baseados em Aprendizagem Profunda 36
3.2.1 WaveNet 36
3.2.2 Tacotron e Tacotron 2 37
3.2.3 FastSpeech 38
3.2.4 Modelos Baseados em Difusão 39

3.2.5 Modelos Recentes de Grande Escala 40
3.2.6 Orpheus TTS e Unsloth 41
3.2.7 Qwen3-TTS 42
3.3 Bases de Dados de Fala para o Português Brasileiro 42
3.3.1 Corpora Públicos 43
3.3.2 Bases Voltadas a Sotaques 44
3.3.3 Limitações Identificadas na Literatura 45
3.4 Síntese da Revisão Bibliográfica 46
3.4.1 Quadro Comparativo dos Trabalhos 46

4 Metodologia **49**
4.1 Visão Geral da Pesquisa 49
4.2 Questões de Pesquisa 50
4.3 Materiais 52
4.3.1 Bases de Dados Utilizadas 52
4.3.2 Modelos de Síntese de Fala Investigados 54
4.3.3 Infraestrutura Computacional 55
4.3.4 Ferramentas e Bibliotecas Utilizadas 55
4.4 Construção do Corpus FALA_PB 56
4.4.1 Extração e Segmentação do CoLingPB 56
4.4.2 Coleta Colaborativa de Falas 57
4.4.3 Filtragem das Bases Sotaque Brasileiro e Braccent 58
4.4.4 Filtragem e Controle de Qualidade 59
4.4.5 Integração de Bases Existentes 60
4.5 Estratégias de Adaptação dos Modelos 61
4.5.1 Clonagem de Fala 62
4.5.2 Ajuste Fino de Modelos Pré-Treinados 63
4.5.3 Treinamento a partir de Pesos Aleatórios 67
4.6 Avaliação Automática dos Modelos 69
4.6.1 Visão Geral e Amostras Avaliadas 69
4.6.2 Preditores de MOS 71

4.6.3 Transcrição Automática das Falas 72
4.6.4 Similaridade Textual: Distância de Levenshtein 72
4.6.5 Métricas de Treinamento 73
4.7 Avaliação Humana do Sotaque 74
4.7.1 Aspectos Éticos 74
4.7.2 Estrutura da Survey 74
4.7.3 Perfil dos Participantes 75
4.7.4 Escalas de Avaliação 75
4.7.5 Estratégias de Comparação entre Áudios 76
4.7.6 Randomização das Amostras 76
4.7.7 Controle de Viés 77
4.7.8 Controle de Fadiga 77
4.7.9 Critérios de Análise dos Resultados 77
4.8 Considerações Éticas 78
4.9 Limitações Metodológicas 79
4.10 Síntese do Capítulo 79

5 Experimentos e Resultados **81**
5.1 Visão Geral dos Experimentos 81
5.2 Caracterização do Corpus FALA_PB 81
5.2.1 Distribuição das Bases Integradas 82
5.2.2 Resultados das Etapas de Filtragem e Segmentação 85
5.2.3 Resultados da Coleta Colaborativa 86
5.2.4 Composição Final do Corpus 86
5.2.5 Limitações Identificadas no Corpus 87
5.2.6 Discussão sobre a Questão de Pesquisa 1 87
5.3 Resultados das Estratégias de Adaptação 89
5.3.1 Resultados da Clonagem de Fala 89
5.3.2 Resultados do Ajuste Fino de Modelos Pré-Treinados 91
5.3.3 Resultados do Treinamento a partir de Pesos Aleatórios 97
5.3.4 Comparação entre Estratégias de Adaptação 99

5.3.5 Condições Experimentais e Proporção de Dados Regionais 101
5.3.6 Discussão sobre a Questão de Pesquisa 2 103
5.4 Resultados das Avaliações Automáticas 105
5.4.1 Métricas e Procedimento de Agregação 106
5.4.2 Inteligibilidade: Similaridade de Levenshtein 107
5.4.3 Qualidade Perceptiva Global: *DNSMOS overall* 108
5.4.4 MOS Principal: *NISQA overall* 109
5.4.5 Dimensões de Qualidade: Ruído ou Mudança de Sotaque? 110
5.4.6 Discussão sobre a Questão de Pesquisa 3 111
5.5 Resultados das Avaliações Subjetivas 114
5.5.1 Estrutura da Avaliação e Perfil dos Participantes 114
5.5.2 Percepção de Sotaque 116
5.5.3 Clareza e Naturalidade: a Adaptação Piora a Qualidade? 117
5.5.4 Comparação Estatística entre os Modelos no Bloco 1 118
5.5.5 Ranqueamento entre os Modelos Adaptados 120
5.5.6 Robustez por Subgrupos 121
5.5.7 Análise Conservadora por Participante 122
5.5.8 Discussão sobre a Questão de Pesquisa 4 122
5.6 Análise Comparativa Geral 123
5.6.1 Considerações Iniciais 123
5.6.2 Convergência entre Avaliação Automática e Subjetiva 124
5.6.3 Ameaças à Validade 125
5.7 Síntese do Capítulo 127

6 Conclusões e Sugestões para Trabalhos Futuros 128
6.1 Conclusões 128
6.2 Sugestões para Trabalhos Futuros 133
6.2.1 Aprofundamento na Caracterização do Sotaque 133
6.2.2 Expansão da Base FALA_PB por Coleta em Redes Sociais 133
6.2.3 Especialização Regional do Sotaque Paraibano 135
6.2.4 Arquiteturas com Modelagem Explícita de Sotaque 135

6.3 Considerações Finais 136

A Prompt para Geração das Mensagens de Coleta Colaborativa 145

A.1 Mensagens Enviadas aos Participantes 146

B Processamento do Corpus Linguístico da Paraíba (CoLingPB) 148

B.1 Extração e Segmentação 148

B.2 Codificação dos Arquivos Originais 149

B.3 Nomenclatura dos Segmentos Processados 151

B.4 Caracterização dos Segmentos do CoLingPB após Filtragem 152

C Conjunto Completo de Gráficos da Avaliação Automática 155

C.1 Inteligibilidade: Similaridade de Levenshtein 156

C.2 MOS Principal: *NISQA overall* 157

C.3 *DNSMOS*: Qualidade Global (*overall*) 159

C.4 *DNSMOS*: Qualidade do Sinal de Voz (*sig*) 160

C.5 *DNSMOS*: Qualidade do Fundo (*bak*) 162

C.6 *NISQA*: Ruído Percebido (*noisiness*) 163

C.7 *NISQA*: Descontinuidade (*discontinuity*) 165

C.8 *NISQA*: Coloração (*coloration*) 166

C.9 *NISQA*: Intensidade Sonora (*loudness*) 168

D Taxas de Erro de Caractere e de Palavra (CER e WER) 170

E Parecer do Comitê de Ética em Pesquisa 174

F Instrumento de Coleta da Avaliação Perceptiva 179

Lista de Abreviaturas e Siglas

ASR - *Automatic Speech Recognition* (Reconhecimento Automático de Fala)

CoLingPB - Corpus Linguístico da Paraíba

DNSMOS - *Deep Noise Suppression Mean Opinion Score*

FALA_PB - Base de dados de fala com sotaque paraibano (corpus desta dissertação)

FFT - *Fast Fourier Transform* (Transformada Rápida de Fourier)

G2P - *Grapheme-to-Phoneme* (Conversão Grafema-Fonema)

GAN - *Generative Adversarial Network* (Rede Generativa Adversária)

GPU - *Graphics Processing Unit* (Unidade de Processamento Gráfico)

HMM - *Hidden Markov Model* (Modelo Oculto de Markov)

ITU-T - *International Telecommunication Union* (Setor de Normalização das Telecomunicações)

LoRA - *Low-Rank Adaptation* (Adaptação de Baixo Posto)

MOS - *Mean Opinion Score* (Pontuação Média de Opinião)

NISQA - *Non-Intrusive Speech Quality Assessment*

PCM - *Pulse-Code Modulation* (Modulação por Código de Pulso)

PEFT - *Parameter-Efficient Fine-Tuning* (Ajuste Fino Eficiente em Parâmetros)

STT - *Speech-to-Text* (Voz para Texto)

TTS - *Text-to-Speech* (Texto para Voz)

VQ-VAE - *Vector Quantized Variational Autoencoder*

WER - *Word Error Rate* (Taxa de Erro de Palavra)

seq2seq - *Sequence-to-Sequence* (Sequência para Sequência)

Lista de Figuras

2.1 Visão geral do processo de síntese de fala: um texto de entrada é processado pelo sintetizador (TTS), que gera como saída um arquivo de áudio contendo a fala correspondente ao texto. 10
2.2 Exemplo simplificado do funcionamento da síntese concatenativa. 12
2.3 Fluxo simplificado de um sistema de síntese paramétrica baseado em HMM. 14
2.4 Arquitetura simplificada de um sistema neural de síntese de fala. 16
2.5 Representação visual de forma de onda de um sinal de fala. 17
2.6 Representação visual de um espectrograma mel de um sinal de fala. 18
2.7 Visão geral do fluxo de um sistema de síntese de fala baseado em aprendizagem profunda, do texto à *waveform*. 20
2.8 Matriz de programação dinâmica da distância de Levenshtein entre a referência “mainha” e a transcrição “minha”. 30

3.1 Arquitetura do Tacotron, ilustrando o fluxo do texto de entrada para a predição do espectrograma mel e posterior geração de *waveform* via *vocoder*. . . 38
3.2 Arquitetura geral do FastSpeech: (a) *Transformer feed-forward*; (b) bloco FFT; (c) regulador de comprimento; (d) preditor de duração. 39
3.3 Visão geral da arquitetura de treinamento do XTTS. 41

4.1 Etapas da metodologia adotada para adaptação de modelos de síntese de fala ao sotaque paraibano. 50
4.2 Fluxo de processamento das bases de fala para construção do corpus FALA_PB. 61

4.3 Fluxo conceitual para estimar a preservação do conteúdo linguístico por meio de transcrição automática e similaridade baseada na distância de Levenshtein. 73

5.1 Histograma de distribuição por estado dos falantes do *dataset* Sotaque Brasileiro. 82
5.2 Histograma de distribuição por sotaque dos falantes do *dataset* Braccent. . . 83
5.3 Curva de perda por passo no ajuste fino do XTTS. 92
5.4 Curva de perda por época de treinamento no ajuste fino do XTTS. 92
5.5 Curva de perda por passo no ajuste fino do F5-TTS. 93
5.6 Curva de perda por época de treinamento no ajuste fino do F5-TTS. 94
5.7 Curva de perda por passo no ajuste fino do Orpheus TTS. 95
5.8 Curva de perda por passo no ajuste fino do Fish Speech. 96
5.9 Curva de perda de treino (*train loss*) no ajuste fino do Qwen3-TTS. 97
5.10 Perda de validação por época no treinamento do F5-TTS a partir de pesos aleatórios (150 épocas). 98
5.11 Similaridade de Levenshtein: distribuição por modelo e condição (azul: original; laranja: ajustado). 108
5.12 *DNSMOS overall*: distribuição por modelo e condição (azul: original; laranja: ajustado). 109
5.13 Caracterização dos participantes da avaliação perceptiva quanto à familiaridade com o sotaque paraibano ($n = 56$). 115
5.14 Distribuição global das respostas por critério de avaliação: escolha de um dos áudios, equivalência percebida (“ambos”) e indistinguibilidade. 116
5.15 Percepção de sotaque paraibano (original $\times$ adaptado) por modelo. A categoria “adaptado tem mais sotaque” (rotulada na legenda) domina amplamente os três modelos. 117
5.16 Clareza e naturalidade percebidas (original $\times$ adaptado) por modelo. Fish Speech e Orpheus TTS apresentam preferência clara pelo adaptado; o XTTS mostra equilíbrio entre as condições. 118

5.17 Taxa de vitória de cada modelo adaptado em sotaque e em clareza/naturalidade (Bloco 2). O XTTS lidera em ambos os critérios, seguido por Fish Speech e Orpheus TTS. 120
5.18 Confrontos diretos par a par entre os modelos adaptados, em sotaque e em clareza/naturalidade. 121

6.1 Fluxo de coleta via *webcrawling* em redes sociais e pré-processamento para extensão da base FALA_PB. 134

B.1 Fluxo de extração, segmentação e filtragem dos áudios do CoLingPB. . . . 149
B.2 Distribuição dos segmentos do CoLingPB processado por zona após filtragem. 152
B.3 Distribuição dos segmentos do CoLingPB processado por sexo do participante após filtragem. 153
B.4 Distribuição dos segmentos do CoLingPB processado por faixa etária após filtragem. 153
B.5 Histograma da duração (em segundos) dos segmentos do CoLingPB processado (após filtragem). 154
B.6 Distribuição do número de palavras por segmento no CoLingPB processado após filtragem. 154

C.1 Similaridade de Levenshtein: médias por modelo e condição. 156
C.2 Similaridade de Levenshtein: distribuição por modelo e condição. 156
C.3 Similaridade de Levenshtein: histograma por modelo e condição. 157
C.4 *NISQA overall* (MOS principal): médias por modelo e condição. 157
C.5 *NISQA overall*: distribuição por modelo e condição. 158
C.6 *NISQA overall*: histograma por modelo e condição. 158
C.7 *DNSMOS overall*: médias por modelo e condição. 159
C.8 *DNSMOS overall*: distribuição por modelo e condição. 159
C.9 *DNSMOS overall*: histograma por modelo e condição. 160
C.10 *DNSMOS sig*: médias por modelo e condição. 160
C.11 *DNSMOS sig*: distribuição por modelo e condição. 161
C.12 *DNSMOS sig*: histograma por modelo e condição. 161
C.13 *DNSMOS bak*: médias por modelo e condição. 162

C.14 *DNSMOS bak*: distribuição por modelo e condição. 162
C.15 *DNSMOS bak*: histograma por modelo e condição. 163
C.16 *NISQA noisiness*: médias por modelo e condição. 163
C.17 *NISQA noisiness*: distribuição por modelo e condição. 164
C.18 *NISQA noisiness*: histograma por modelo e condição. 164
C.19 *NISQA discontinuity*: médias por modelo e condição. 165
C.20 *NISQA discontinuity*: distribuição por modelo e condição. 165
C.21 *NISQA discontinuity*: histograma por modelo e condição. 166
C.22 *NISQA coloration*: médias por modelo e condição. 166
C.23 *NISQA coloration*: distribuição por modelo e condição. 167
C.24 *NISQA coloration*: histograma por modelo e condição. 167
C.25 *NISQA loudness*: médias por modelo e condição. 168
C.26 *NISQA loudness*: distribuição por modelo e condição. 168
C.27 *NISQA loudness*: histograma por modelo e condição. 169

D.1 CER e WER médios por modelo e condição. 172
D.2 CER: distribuição por modelo e condição. 173
D.3 WER: distribuição por modelo e condição. 173

Lista de Tabelas

2.1 Exemplo de escala de *Mean Opinion Score* (MOS) para avaliação de inteligibilidade da fala sintetizada. 27

3.1 Número de horas de áudio por idioma no conjunto de dados utilizado no treinamento do XTTS. 44
3.2 Quadro comparativo de trabalhos relacionados à síntese de fala, corpora linguísticos e modelagem de sotaques. 46

4.1 Infraestrutura computacional utilizada nos experimentos. 55
4.2 Hiperparâmetros do ajuste fino do F5-TTS. 65
4.3 Hiperparâmetros do ajuste fino do Orpheus TTS. 66
4.4 Hiperparâmetros do ajuste fino do Fish Speech. 66
4.5 Hiperparâmetros do ajuste fino do Qwen3-TTS. 67
4.6 Hiperparâmetros do treinamento do F5-TTS a partir de pesos aleatórios. . . 68
4.7 Métricas automáticas calculadas sobre cada amostra de fala sintetizada. . . 70
4.8 Comparação entre diferentes metodologias de avaliação em síntese de fala. . 71
4.9 Exemplos de similaridade baseada na distância de Levenshtein entre o texto original e o texto reconhecido por STT. 73

5.1 Distribuição de entrevistas do CoLingPB por município. 84
5.2 Métricas do CoLingPB após extração, segmentação e filtragem. 85
5.3 Composição final das amostras da coleta colaborativa. 86
5.4 Resumo das bases de dados de fala utilizadas na construção da base FALA_PB. 87
5.5 Resumo do treinamento do modelo XTTS. 91
5.6 Resumo do treinamento do F5-TTS. 93

5.7 Resumo do treinamento do Orpheus TTS. 95
5.8 Resumo do treinamento do Fish Speech. 95
5.9 Resumo do treinamento do Qwen3-TTS. 96
5.10 Resumo da execução do treinamento do F5-TTS a partir de pesos aleatórios. 98
5.11 Comparação das estratégias de adaptação quanto ao tempo de computação e ao desfecho qualitativo preliminar. 100
5.12 Condições experimentais e desfecho por experimento. 101
5.13 Similaridade de Levenshtein (inteligibilidade) por modelo: original vs. ajustado. 107
5.14 *DNSMOS overall* (qualidade perceptiva global) por modelo: original vs. ajustado. 108
5.15 *NISQA overall* (MOS principal) por modelo: original vs. ajustado. 109
5.16 Dimensões de qualidade (*NISQA* e *DNSMOS*) por modelo: $\Delta = \overline{\text{ajustado}} - \overline{\text{original}}$ (*: $p < 0{,}05$). 110
5.17 Percepção de sotaque paraibano (Bloco 1, todos os respondentes, $n = 56$; 168 respostas por modelo). 117
5.18 Clareza e naturalidade percebidas (Bloco 1, todos os respondentes, $n = 56$; 168 respostas por modelo). 118
5.19 Comparações par a par entre os modelos (Bloco 1): teste exato de Fisher com correção de Holm sobre as taxas favoráveis. 119
5.20 Ranqueamento dos modelos adaptados por taxa de vitória (Bloco 2, todos os respondentes; empates excluídos). 120
5.21 Convergência entre as avaliações automática e subjetiva. 125

B.1 Partes do padrão de nomenclatura dos arquivos originais do CoLingPB. . . 150

D.1 Similaridade de Levenshtein, CER e WER por modelo. 171

Capítulo 1

Introdução

A síntese de fala evoluiu de métodos concatenativos e baseados em regras para arquiteturas neurais profundas capazes de gerar vozes altamente naturais e inteligíveis. Apesar desses avanços, os sistemas abertos de síntese de fala ainda representam de forma limitada a diversidade de sotaques regionais, sobretudo variedades brasileiras pouco presentes nos corpora públicos, como o sotaque paraibano. Este capítulo contextualiza o tema, apresenta as motivações da pesquisa, delimita o problema investigado, enuncia as questões de pesquisa e as hipóteses e descreve os objetivos, as contribuições e a organização da dissertação.

1.1 Contextualização

A síntese de fala era tradicionalmente realizada por meio de métodos concatenativos e sistemas baseados em regras, nos quais segmentos de áudio previamente gravados eram organizados de maneira sequencial para formar a fala sintetizada (Hunt e Black, 1996; Zen et al., 2009). Apesar dos avanços obtidos por essas abordagens, os sistemas produzidos frequentemente apresentavam limitações relacionadas à naturalidade, fluidez e expressividade da fala.

Com o avanço das técnicas de aprendizado de máquina e, posteriormente, das redes neurais profundas, a área de síntese de fala passou por mudanças significativas. Modelos neurais modernos passaram a permitir uma modelagem mais precisa das características acústicas e prosódicas da fala, resultando em vozes sintetizadas mais naturais e inteligíveis (Shen et al., 2018; van den Oord et al., 2016).

Atualmente, o estado da arte em síntese de fala é dominado por arquiteturas neurais profundas capazes de converter texto em fala com elevados níveis de qualidade perceptual. Esses sistemas vêm sendo aplicados em diferentes contextos, incluindo assistentes virtuais (Hoy, 2018), interfaces conversacionais (Cambre e Kulkarni, 2019), recursos de acessibilidade para pessoas com deficiência visual (Freitas e Kouroupetroglou, 2008), além de narração automatizada de audiolivros (Zhang et al., 2023) e produção de conteúdo audiovisual, como a dublagem automática (Federico et al., 2020).

Apesar desses avanços, os sistemas de síntese de fala apresentam limitações para representar a variação regional da fala. Os modelos e corpora públicos concentram-se em variedades linguísticas predominantes, o que restringe a capacidade desses sistemas de reproduzir sotaques e regionalismos menos representados. Embora estudos como o de Batista (2019) demonstrem a relevância das variações regionais da fala brasileira, ainda existe uma quantidade limitada de pesquisas voltadas à incorporação desses regionalismos em sistemas abertos de síntese de fala com aprendizagem profunda, particularmente no contexto do sotaque paraibano.

1.2 Motivações

A representação de sotaques regionais em sistemas de síntese de fala possui relevância não apenas técnica, mas também social e cultural. A capacidade de incorporar características regionais da fala em interfaces computacionais pode contribuir para experiências mais naturais, acessíveis e culturalmente representativas para diferentes grupos de usuários (Cambre e Kulkarni, 2019; Tinchev et al., 2023).

No contexto das interações humano-computador, sistemas capazes de refletir nuances linguísticas e culturais tendem a promover interações mais inclusivas e próximas da realidade dos usuários. Além disso, aplicações como narração automatizada, audiolivros, jogos digitais, produções audiovisuais e prototipação de conteúdo multimídia podem se beneficiar diretamente da utilização de vozes sintetizadas com características regionais específicas.

Entretanto, muitas das soluções mais avançadas atualmente disponíveis em síntese de fala encontram-se restritas a plataformas proprietárias e fechadas, como ElevenLabs (ElevenLabs, 2026), Amazon Polly (Amazon Web Services, 2026), Google Cloud Text-to-Speech

(Google Cloud, 2026) e Microsoft Azure AI Speech (Microsoft, 2026). Ainda que esses serviços comerciais reportem alta naturalidade e expressividade em seus relatórios técnicos (Atamanenko et al., 2025), o acesso restrito aos modelos e às metodologias impõe limitações relevantes. Do ponto de vista da privacidade, os textos de entrada precisam ser enviados a servidores de terceiros, o que pode ser inadequado para dados sensíveis. Quanto ao custo, o uso costuma ser tarifado por volume de áudio gerado, encarecendo aplicações de larga escala ou de uso contínuo. Em relação à flexibilidade de configuração, o usuário fica restrito às opções expostas pela interface, sem controle sobre a arquitetura ou os dados de treinamento. Por fim, a impossibilidade de inspecionar pesos, dados e procedimentos compromete a reprodutibilidade científica e dificulta adaptações a contextos linguísticos específicos, como os sotaques regionais. Nesse cenário, torna-se relevante investigar abordagens abertas e reprodutíveis para síntese de fala com foco em sotaques regionais brasileiros, em especial o sotaque paraibano, característico da Paraíba, Estado da Federação onde esta pesquisa se realiza, cuja escassez de recursos e estudos evidencia uma lacuna na literatura.

1.3 Problema de Pesquisa

Apesar dos avanços recentes em modelos neurais de síntese de fala, persiste uma baixa representatividade de sotaques regionais brasileiros em sistemas abertos. Essa constatação apoia-se no levantamento de bases de dados e trabalhos relacionados apresentado no Capítulo 3, que evidencia a concentração dos recursos públicos de fala em variantes mais predominantes, sobretudo das regiões Sudeste e Sul, e a escassez de estudos voltados a variedades do Nordeste (Batista, 2019; Tinchev et al., 2023). No caso do sotaque paraibano, essa lacuna se manifesta tanto na ausência de bases de dados representativas quanto na falta de estratégias de adaptação capazes de preservar naturalidade e inteligibilidade ao incorporar as características regionais da fala.

Dessa forma, esta pesquisa investiga como modelos abertos de síntese de fala podem ser adaptados para incorporar características do sotaque paraibano e como a qualidade dessas adaptações pode ser avaliada de forma objetiva, por métricas automáticas de qualidade e inteligibilidade, e subjetiva, por avaliação perceptiva conduzida com ouvintes humanos. Para viabilizar e mensurar essas adaptações, a pesquisa também aborda a construção de um corpus

de fala com sotaque paraibano e de um protocolo de avaliação reprodutível, recursos sem os quais tais adaptações não poderiam ser realizadas nem comparadas de maneira sistemática.

Para fins de escopo, esta dissertação delimita a investigação a três estratégias de adaptação de modelos abertos de síntese de fala ao sotaque paraibano, em ordem crescente de esforço computacional e de dependência de dados: clonagem de voz, ajuste fino de modelos pré-treinados e treinamento a partir de pesos aleatórios. Essas três técnicas constituem o eixo comparativo do trabalho, e o desenho experimental, os modelos selecionados e os protocolos de avaliação são organizados em torno delas.

1.4 Questões de Pesquisa

A partir do problema delimitado, esta pesquisa é orientada por quatro questões de pesquisa, apresentadas a seguir de forma sintética e detalhadas, com seus procedimentos associados, no Capítulo 4:

- **QP1:** Como construir uma base de dados de fala voltada ao sotaque paraibano adequada para o treinamento e a adaptação de modelos de síntese de fala?
- **QP2:** Quais estratégias de adaptação são viáveis para a incorporação do sotaque paraibano em modelos de síntese de fala, considerando as restrições de recursos computacionais e de dados disponíveis?
- **QP3:** Os ajustes realizados nos modelos de síntese de fala impactam negativamente a qualidade e a inteligibilidade das falas sintetizadas?
- **QP4:** Os ajustes realizados nos modelos de síntese de fala conseguem produzir percepção consistente do sotaque paraibano nas falas sintetizadas?

1.5 Hipóteses

A cada questão de pesquisa associa-se uma hipótese de trabalho, enunciada a seguir de forma sintética e detalhada, com os procedimentos de verificação, no Capítulo 4. Como as adaptações são comparadas por meio de avaliação estatística das métricas automáticas e percepti-

vas, as hipóteses são operacionalizadas nos experimentos e retomadas, à luz das evidências obtidas, no Capítulo 5 e nas conclusões do Capítulo 6:

- **H1** (associada à QP1): A integração de bases públicas existentes, o pré-processamento de corpora linguísticos e a coleta colaborativa complementar permitem construir um corpus com diversidade e qualidade suficientes para adaptar modelos de síntese de fala ao sotaque paraibano.
- **H2** (associada à QP2): As três estratégias de adaptação (clonagem de voz, ajuste fino e treinamento a partir de pesos aleatórios), por diferirem em esforço computacional e volume de dados, apresentam capacidades distintas de reproduzir o sotaque paraibano.
- **H3** (associada à QP3): A incorporação do sotaque paraibano pela melhor técnica identificada não degrada de forma estatisticamente significativa a qualidade e a inteligibilidade das falas sintetizadas em relação aos modelos originais, o que é verificável por preditores de MOS e por métricas de similaridade textual sobre transcrições automáticas.
- **H4** (associada à QP4): Os modelos adaptados sobre o corpus *FALA_PB* produzem percepção consistente e estatisticamente significativa do sotaque paraibano junto a ouvintes humanos, sem prejuízo da clareza e da naturalidade percebidas.

1.6 Objetivos

1.6.1 Objetivo Geral

Investigar abordagens abertas de síntese de fala com aprendizagem profunda neural com foco na incorporação de características do sotaque paraibano em sistemas de síntese de fala para o português brasileiro.

1.6.2 Objetivos Específicos

- comparar diferentes estratégias de adaptação de modelos neurais de síntese de fala (clonagem de voz, ajuste fino e treinamento a partir de pesos aleatórios) aplicadas ao contexto do sotaque paraibano;

- construir uma base de dados voltada à representação do sotaque paraibano a partir da integração, processamento e complementação de bases públicas existentes;
- avaliar a capacidade de modelos abertos de síntese de fala em reproduzir características regionais da fala paraibana;
- investigar métricas e estratégias de avaliação relacionadas à naturalidade, inteligibilidade e adequação regional das vozes sintetizadas.

1.7 Contribuições

As principais contribuições desta dissertação, além das abordagens metodológicas próprias e das análises de seus resultados, são recursos abertos e reprodutíveis voltados à síntese de fala com sotaque paraibano no contexto do português brasileiro. Entre os resultados alcançados pelo trabalho, destacam-se:

- o corpus *FALA_PB*, base de dados de fala com sotaque paraibano construída a partir da integração de bases públicas e de uma coleta colaborativa complementar com falantes da Paraíba, disponibilizado publicamente na plataforma Zenodo (de Morais, 2026);
- *checkpoints* de modelos abertos de síntese de fala adaptados ao sotaque paraibano, disponibilizados para uso e reprodução na plataforma Hugging Face[1];
- o código do *pipeline* de construção do corpus, adaptação dos modelos e avaliação, disponibilizado em repositório aberto sob licença MIT[2];
- um protocolo de avaliação que combina métricas objetivas de qualidade e inteligibilidade com avaliação perceptiva conduzida com ouvintes humanos;
- evidências comparativas sobre a adequação de diferentes estratégias de adaptação para a incorporação de sotaque regional em modelos de síntese de fala.

Busca-se, com isso, ampliar a representatividade regional e disponibilizar recursos que possam ser utilizados em pesquisas futuras sobre sotaques brasileiros, síntese de fala e tecnologias abertas de fala.

[1] `https://huggingface.co/acidente/fala_pb_checkpoints`
[2] `https://github.com/ThomazDiniz/fala_pb_scripts`

1.8 Declaração de Uso de Inteligência Artificial Generativa

Em atenção às boas práticas de transparência na pesquisa científica, declara-se o uso de ferramentas de inteligência artificial generativa como apoio a etapas específicas do desenvolvimento deste trabalho. Não estão abrangidos por esta declaração os modelos de síntese de fala investigados (XTTS, F5-TTS, Orpheus TTS, Fish Speech e Qwen3-TTS), uma vez que constituem o próprio objeto de estudo da pesquisa e são descritos em detalhe na Seção 4.3.2.

O emprego dessas ferramentas restringiu-se às seguintes etapas:

- **Suporte à implementação.** Assistentes de programação foram utilizados no apoio à configuração de projetos e de ambientes de execução em contêineres *Docker*, empregados para isolar as dependências de cada modelo de síntese de fala investigado. Foram utilizados o Claude, da Anthropic (modelos Sonnet, Fable 5, Opus 4.5 e Opus 5), e o editor Cursor (modelo GPT-4o, da OpenAI);

- **Geração de gráficos.** As mesmas ferramentas apoiaram a escrita dos *scripts* de visualização utilizados na apresentação dos resultados das avaliações automática e perceptiva (Capítulo 5 e Apêndice C);

- **Preparação de dados da coleta colaborativa.** As frases lidas pelos participantes voluntários foram geradas de forma assistida pelo ChatGPT, da OpenAI (modelo GPT-4o), a partir do *prompt* estruturado reproduzido no Apêndice A e descrito na Seção 4.4.2;

- **Revisão linguística e de estilo.** O texto da dissertação passou por revisão de português assistida pelo Claude, da Anthropic, abrangendo ortografia, pontuação, concordância e padronização terminológica.

Em todos os casos, o uso limitou-se a apoio operacional e instrumental: as saídas produzidas por essas ferramentas foram integralmente revisadas, testadas e validadas pelo autor antes de sua incorporação à pesquisa. A concepção do problema, o desenho experimental, a construção e a curadoria do corpus, a condução dos experimentos, a análise e a interpretação dos resultados não foram delegadas a ferramentas de inteligência artificial generativa.

Declara-se, por fim, que a autoria intelectual, a curadoria e a responsabilidade integral pelo conteúdo final desta pesquisa, incluindo as implementações, os dados, os resultados e o próprio texto da dissertação, pertencem exclusivamente ao seu autor humano.

1.9 Organização da Dissertação

A organização do restante desta dissertação contempla os seguintes capítulos. No Capítulo 2 é apresentada a fundamentação teórica relacionada à síntese de fala, aos modelos de aprendizado profundo aplicados à fala, aos sotaques e variações linguísticas no português brasileiro e às abordagens de avaliação de sistemas de síntese de fala. No Capítulo 3 são apresentados a revisão da literatura e os trabalhos relacionados, situando esta pesquisa em relação a estudos anteriores sobre síntese de fala, adaptação de modelos e representação de sotaques. No Capítulo 4 é descrita a metodologia da pesquisa, incluindo a construção e o processamento do corpus *FALA_PB*, os modelos selecionados, as estratégias de adaptação investigadas e os protocolos de avaliação automática e manual. No Capítulo 5 são apresentados os experimentos realizados e os resultados obtidos, percorrendo todo o fluxo da pesquisa, desde a adaptação dos modelos por meio do treinamento com o corpus *FALA_PB* até a avaliação das vozes geradas, de forma automática (métricas objetivas de qualidade e inteligibilidade) e manual (percepção de ouvintes quanto à qualidade e à representação do sotaque paraibano). Por fim, no Capítulo 6 são apresentadas as considerações finais da pesquisa, suas limitações e possíveis direções para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo reúne os conceitos necessários à compreensão da pesquisa. Inicialmente, apresenta-se o conceito de síntese de fala e sua evolução histórica, dos sistemas concatenativos aos modelos neurais atuais. Em seguida, são revistos os fundamentos de redes neurais artificiais e as representações de áudio empregadas por esses modelos, seguidos da descrição dos componentes de um sistema de síntese de fala e das estratégias de reaproveitamento de modelos (pré-treinamento, ajuste fino e clonagem de voz). Depois, discutem-se os sotaques e a variação linguística no português brasileiro, base sociolinguística do trabalho. Por fim, apresentam-se as abordagens de avaliação de sistemas de síntese de fala, incluindo métricas subjetivas, preditores automáticos de qualidade e testes de significância estatística.

2.1 Síntese de Fala

A síntese de fala, também conhecida como *Text-to-Speech* (TTS), consiste no processo de geração artificial de fala a partir de uma entrada textual. Sistemas de síntese de fala têm como objetivo transformar texto escrito em sinais de áudio inteligíveis, buscando reproduzir características naturais da fala humana, como pronúncia, ritmo, entonação e prosódia, conforme define Dutoit (1997) em sua introdução clássica à área. A Figura 2.1 ilustra esse processo de forma esquemática: o texto de entrada é processado pelo sintetizador, que produz como saída um arquivo de áudio contendo a fala correspondente.

Figura 2.1: Visão geral do processo de síntese de fala: um texto de entrada é processado pelo sintetizador (TTS), que gera como saída um arquivo de áudio contendo a fala correspondente ao texto.



Fonte: elaborado pelo autor (2026).

Atualmente, sistemas TTS são utilizados em diversas aplicações, incluindo assistentes virtuais, leitores de tela, audiolivros, sistemas de navegação, interfaces conversacionais, jogos digitais e ferramentas de acessibilidade. Nessas aplicações, a qualidade perceptiva da fala sintetizada exerce impacto direto na experiência do usuário, tornando relevantes aspectos como inteligibilidade, naturalidade e adaptação linguística (Zen et al., 2009; Tan et al., 2021).

2.1.1 Evolução Histórica da Síntese de Fala

A síntese de fala evoluiu de sistemas baseados em regras linguísticas e modelos articulatórios, revisados por Klatt (1987), passando pela síntese concatenativa por seleção de unidades de Hunt e Black (1996) e pela síntese paramétrica estatística baseada em modelos ocultos de Markov (*Hidden Markov Models* ou HMMs), consolidada por Zen et al. (2009), até as abordagens neurais profundas que dominam o estado da arte atual (van den Oord et al., 2016; Cui et al., 2025). Cada paradigma buscou superar as limitações do anterior em naturalidade, flexibilidade e custo computacional, e é detalhado nas subseções seguintes.

2.1.2 Síntese Concatenativa

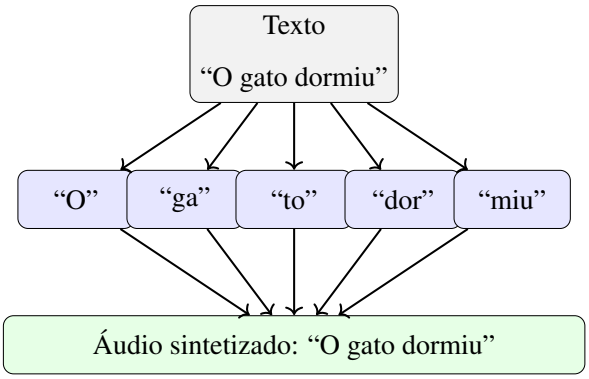
A síntese concatenativa consiste em uma abordagem de síntese de fala baseada na concatenação de segmentos de áudio previamente gravados. Nesse paradigma, grandes bases de dados de fala são utilizadas para armazenar unidades linguísticas, como fonemas, sílabas, palavras ou trechos maiores de fala, que posteriormente são selecionadas e concatenadas para formar a sentença desejada, conforme o paradigma de seleção de unidades proposto por Hunt e Black (1996).

Durante o processo de síntese, o sistema realiza a seleção das unidades acústicas consideradas mais adequadas ao contexto linguístico e prosódico do texto de entrada. Em seguida, essas unidades são concatenadas para produzir o sinal de áudio final. A qualidade perceptiva da fala sintetizada depende diretamente da cobertura e diversidade da base de dados utilizada, bem como da eficiência dos algoritmos de seleção e junção acústica (Dutoit, 1997).

Uma das principais vantagens da síntese concatenativa está na elevada naturalidade da fala gerada, especialmente quando comparada às abordagens paramétricas clássicas, já que os segmentos utilizados correspondem a gravações reais da voz humana. Em contrapartida, esses sistemas apresentam limitações importantes: forte dependência de grandes bases de dados (seriam necessárias centenas de horas de gravação para reproduzir adequadamente um novo locutor), baixa flexibilidade para modificar prosódia ou sotaques específicos e descontinuidades acústicas entre segmentos, que podem introduzir artefatos perceptíveis quando as unidades selecionadas diferem em entonação, intensidade ou contexto fonético (Zen et al., 2009). Essas limitações motivaram o desenvolvimento de abordagens mais flexíveis, baseadas em modelagem estatística e em redes neurais profundas.

A Figura 2.2 ilustra a seleção e a concatenação de segmentos previamente gravados para formar a sentença final.

Figura 2.2: Exemplo simplificado do funcionamento da síntese concatenativa.



Segmentos de fala previamente gravados são selecionados e concatenados para formar a sentença final.

Fonte: elaborado pelo autor (2026).

2.1.3 Síntese Paramétrica

A síntese paramétrica surgiu como alternativa às limitações observadas nos sistemas concatenativos. Diferentemente das abordagens baseadas em gravações completas de unidades acústicas, os sistemas paramétricos representam a fala por meio de parâmetros matemáticos extraídos do sinal de áudio, permitindo maior flexibilidade e controle sobre a síntese, como sistematizam Zen et al. (2009) em sua revisão sobre o paradigma.

Nessa abordagem, características acústicas da fala, como frequência fundamental, espectro e duração fonética, são modeladas estatisticamente a partir de corpora de treinamento. Durante a síntese, esses parâmetros são utilizados para reconstruir artificialmente o sinal de áudio utilizando modelos vocais ou vocoders.

Modelos ocultos de Markov (HMMs) tornaram-se amplamente utilizados em sistemas paramétricos devido à sua capacidade de modelar sequências temporais e variabilidade acústica da fala (Zen et al., 2009; Tokuda et al., 2013). Esses sistemas permitiram maior generalização e adaptação para diferentes vozes e estilos de fala, além de reduzirem a dependência de bases extremamente extensas.

Sistemas paramétricos baseados em HMM foram amplamente empregados em aplicações comerciais de síntese de fala, especialmente em sistemas embarcados, navegação por voz e assistentes virtuais devido ao menor custo computacional e maior flexibilidade em relação aos sistemas concatenativos (Tokuda et al., 2013).

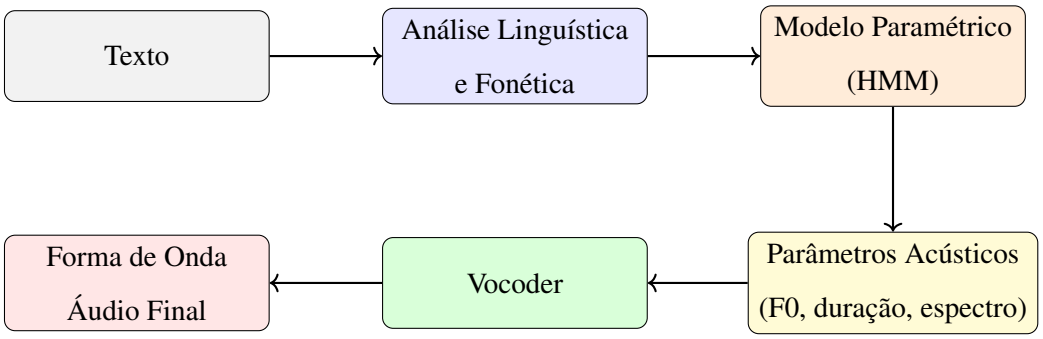
Entretanto, apesar das vantagens relacionadas à flexibilidade e controle, a síntese paramétrica frequentemente produzia fala considerada excessivamente artificial ou robótica quando comparada à fala humana natural. Como apontam van den Oord et al. (2016), esse fenômeno ocorria principalmente devido à suavização excessiva dos parâmetros acústicos durante o processo de modelagem estatística.

As limitações perceptivas da síntese paramétrica motivaram o desenvolvimento de arquiteturas neurais capazes de modelar diretamente distribuições acústicas complexas da fala humana, impulsionando o avanço da síntese neural.

A Figura 2.3 mostra como sistemas de síntese paramétrica utilizam modelos estatísticos para gerar parâmetros acústicos, posteriormente convertidos em áudio por um vocoder.

Os três parâmetros acústicos indicados na figura correspondem às dimensões usualmente modeladas nesse paradigma. A frequência fundamental, denotada por *F0* e medida em hertz, corresponde à taxa de vibração das pregas vocais nos segmentos sonoros e constitui o correlato acústico da altura percebida da voz; é dela que dependem os contornos de entonação e boa parte da expressividade da fala (Cagliari, 1992; Moraes, 2008). A duração corresponde ao tempo de realização de cada unidade fonética e determina o ritmo, a velocidade de elocução e o padrão de acentuação do enunciado (Ladefoged e Johnson, 2012). O espectro, por sua vez, descreve a distribuição de energia do sinal ao longo das frequências em cada instante e carrega a informação que distingue um som de outro, incluindo o timbre da voz e a identidade dos segmentos vocálicos e consonantais; em sistemas paramétricos, ele é tipicamente representado de forma compacta por coeficientes cepstrais em escala mel, conforme detalhado na Seção 2.3.1 (Logan, 2000; Rabiner e Juang, 1993). O modelo estatístico prevê esses parâmetros a partir do texto analisado, e o *vocoder* os utiliza para reconstruir a forma de onda final.

Figura 2.3: Fluxo simplificado do funcionamento de um sistema de síntese paramétrica baseado em HMM.



Fonte: elaborado pelo autor (2026).

2.1.4 Síntese por Aprendizagem Profunda

Atualmente, modelos de aprendizagem profunda representam o estado da arte em síntese de fala. Diferentemente das abordagens concatenativas e paramétricas tradicionais, os sistemas neurais utilizam redes neurais profundas para aprender diretamente as relações entre texto e representações acústicas da fala (van den Oord et al., 2016).

Essa transição teve marcos como o WaveNet (van den Oord et al., 2016), que gera diretamente a forma de onda do áudio, e arquiteturas ponta a ponta como o Tacotron (Shen et al., 2018), que produzem espectrogramas mel a partir do texto, posteriormente convertidos em waveform por vocoders neurais. Mais recentemente, modelos não autorregressivos e baseados em *Transformers* e em processos de difusão, como FastSpeech (Ren et al., 2019), Grad-TTS (Popov et al., 2021) e NaturalSpeech 3 (Ju et al., 2024), passaram a equilibrar qualidade perceptiva, eficiência computacional e controle sobre os atributos da fala. Essas arquiteturas são detalhadas na revisão da literatura (Capítulo 3).

Além da melhoria na naturalidade da síntese, sistemas neurais modernos também possibilitam tarefas como clonagem de voz, síntese multilíngue, adaptação para novos locutores e incorporação de sotaques regionais. Essas capacidades tornam a síntese neural particularmente relevante para aplicações relacionadas à acessibilidade, interfaces conversacionais e preservação de características linguísticas regionais.

Um desdobramento relevante dessa evolução é a chamada síntese controlável, na qual o objetivo deixa de ser apenas produzir fala natural e passa a incluir o controle explícito de

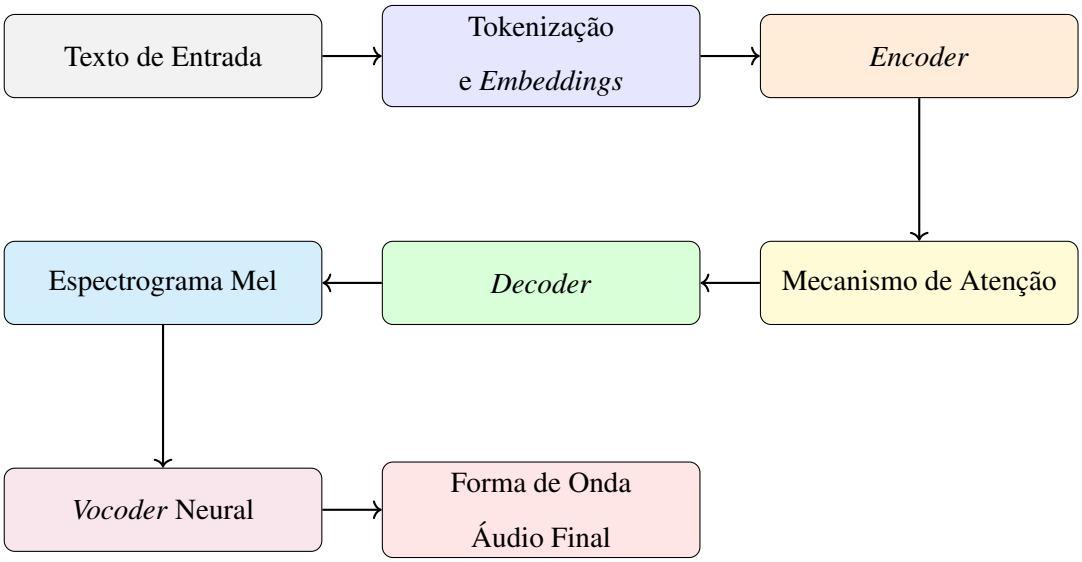
atributos como emoção, timbre, estilo e sotaque. Em levantamento sistemático sobre o tema, Xie et al. (2025) organizam as abordagens existentes em duas grandes famílias: as técnicas tradicionais de controle, baseadas em rótulos discretos ou em representações intermediárias explícitas (identificadores de locutor, marcadores de estilo, curvas de prosódia), e as abordagens mais recentes, em que o controle é expresso em linguagem natural por meio de *prompts* interpretados por modelos de linguagem de grande escala. Os autores destacam ainda que a avaliação da controlabilidade permanece um problema em aberto, uma vez que as métricas usuais de qualidade e inteligibilidade não medem, por si sós, se o atributo pretendido foi de fato realizado na fala gerada.

No que diz respeito especificamente ao sotaque, essa literatura oferece dois caminhos alternativos ao adotado nesta dissertação. O primeiro é o ajuste fino, estratégia investigada neste trabalho, em que os pesos de um modelo pré-treinado são atualizados sobre um corpus da variedade linguística alvo, de modo que o sotaque passa a estar embutido no próprio *checkpoint* resultante (Seção 2.6.2). O segundo consiste em modelar o sotaque como um atributo condicionante, mantendo-o separado da identidade do locutor. É o caso do CrossAccent-TTS, proposto por Annamdevula et al. (2026), que combina representações desemaranhadas de locutor e de sotaque com um controlador de intensidade: durante a inferência, *embeddings* de idioma são combinados de forma ponderada, permitindo interpolar entre sotaques e ajustar continuamente a intensidade do sotaque sem alterar o timbre do falante.

A distinção entre os dois caminhos é metodologicamente importante para esta dissertação. A modelagem por atributo condicionante oferece controle mais fino e dispensa um novo treinamento a cada variedade linguística, mas exige corpora rotulados por sotaque e arquiteturas projetadas para essa separação, condições não satisfeitas pelos modelos abertos e pelo corpus disponíveis neste trabalho. O ajuste fino, em contrapartida, opera sobre arquiteturas já existentes e sobre um corpus regional de porte modesto, ao custo de produzir um modelo por variedade e de não oferecer controle gradual sobre a intensidade do sotaque. Essa limitação é retomada nas discussões dos resultados e nas propostas de trabalhos futuros (Capítulo 6).

A Figura 2.4 apresenta os múltiplos componentes sucessivos de uma solução fundamentada em redes neurais profundas, os quais transformam texto em representações acústicas e, em seguida, em áudio sintetizado.

Figura 2.4: Fluxo simplificado de uma arquitetura neural de síntese de fala, inspirada em modelos como Tacotron e FastSpeech.



Fonte: elaborado pelo autor (2026).

2.2 Redes Neurais Artificiais

Os modelos de síntese de fala investigados nesta dissertação são, em essência, redes neurais artificiais de grande escala (também conhecidas como redes neurais profundas). Esta seção apresenta os fundamentos dessas redes e a arquitetura *Transformer*, base dos sistemas TTS modernos, fornecendo o vocabulário conceitual para as seções seguintes, nas quais os termos pesos, camadas e atenção são utilizados recorrentemente.

2.2.1 Redes Densas e o Conceito de Pesos

Uma rede neural artificial é composta por camadas de unidades que aplicam transformações lineares seguidas de funções de ativação não lineares; o empilhamento dessas camadas produz modelos capazes de aproximar funções complexas, como o mapeamento entre texto e representações acústicas da fala (Goodfellow et al., 2016). O conjunto de matrizes de pesos e vieses dessas camadas constitui os parâmetros do modelo, referidos ao longo deste capítulo simplesmente como pesos: são eles que o pré-treinamento aprende, o ajuste fino modifica e a clonagem *zero-shot* mantém congelados (Seções 2.6.1 a 2.6.3).

2.2.2 *Transformers* e o Mecanismo de Atenção

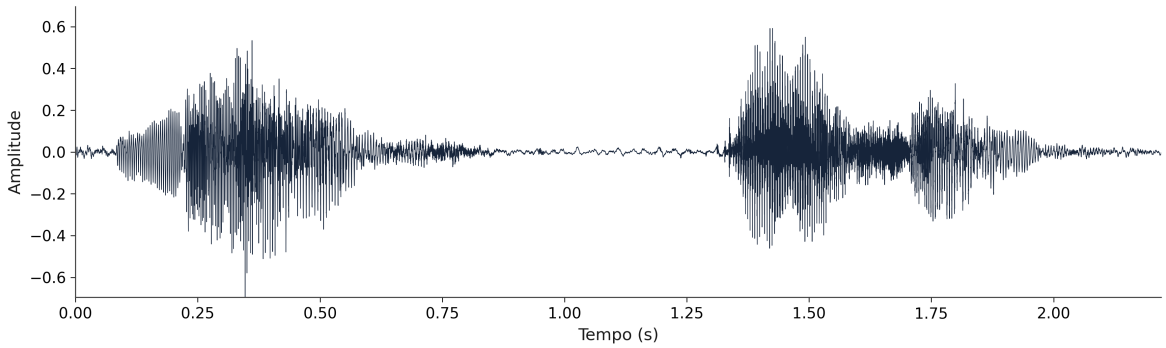
A arquitetura *Transformer*, proposta por Vaswani et al. (2017), é a base da maioria dos sistemas TTS modernos. Seu componente central é o mecanismo de atenção: ao processar uma sequência (de texto ou de áudio), cada elemento consulta todos os demais e aprende a decidir quais são mais relevantes para compor sua própria representação. Em síntese de fala, esse acesso direto a qualquer posição da sequência facilita o alinhamento entre texto e áudio e a captura de dependências prosódicas de longo alcance, além de permitir treinamento altamente paralelizável, ao contrário das redes recorrentes que o antecederam. Arquiteturas como FastSpeech (Ren et al., 2019) e XTTS (Casanova et al., 2024) são baseadas em *Transformers*, e a atenção é também o mecanismo que permite condicionar a síntese a um áudio de referência na clonagem de voz (Seção 2.6.3).

2.3 Representações de Áudio

2.3.1 Forma de Onda, Espectrograma e Escala Mel

O sinal de fala pode ser representado digitalmente pela forma de onda (*waveform*), que descreve a variação de amplitude ao longo do tempo (Figura 2.5). Em síntese neural, a forma de onda costuma ser a saída final do sistema, mas sua elevada dimensionalidade leva muitos modelos a operar sobre representações intermediárias mais compactas (van den Oord et al., 2016).

Figura 2.5: Representação visual de forma de onda de um sinal de fala.

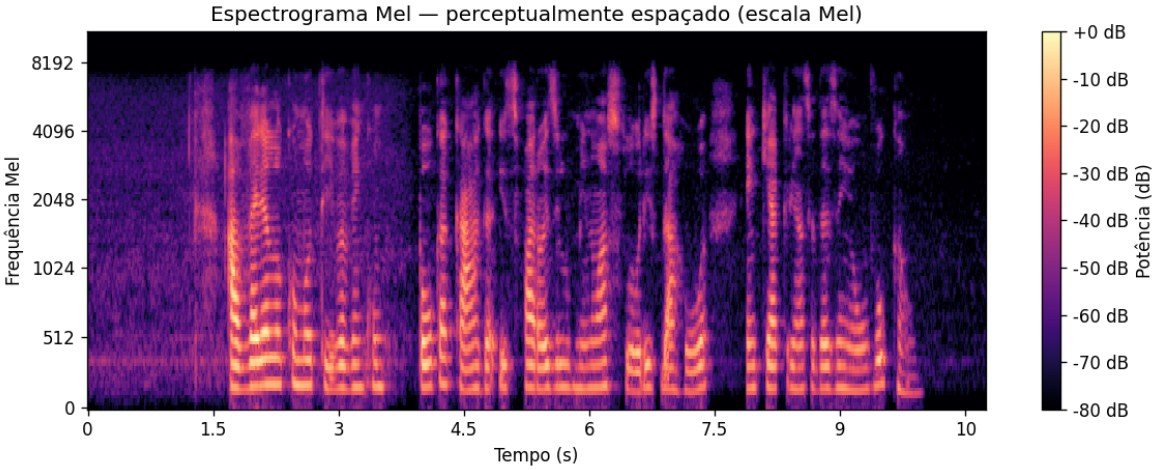


Fonte: elaborado pelo autor (2026).

Muitas características da fala estão associadas à distribuição de frequências do sinal,

analisada no domínio espectral por meio da Transformada de Fourier (Smith, 1997). O espectrograma representa essa distribuição espectral ao longo do tempo, evidenciando padrões fonéticos e prosódicos como formantes e transições (Rabiner e Juang, 1993). O espectrograma Mel é uma variante baseada na escala Mel, perceptualmente inspirada na audição humana, que dá maior resolução às frequências baixas e reduz a dimensionalidade preservando informação relevante para a fala (Stevens et al., 1937; Logan, 2000) (Figura 2.6). Por essas propriedades, o espectrograma Mel é a representação intermediária mais comum em modelos neurais como Tacotron e FastSpeech, posteriormente convertida em forma de onda por vocoders (Shen et al., 2018; Ren et al., 2019).

Figura 2.6: Representação visual de um espectrograma mel de um sinal de fala.



Fonte: elaborado pelo autor (2026).

A escala Mel é uma escala perceptual de altura, obtida experimentalmente a partir de julgamentos de ouvintes (Stevens et al., 1937), sendo aproximadamente linear nas frequências baixas e progressivamente comprimida nas altas. No âmbito da síntese de fala, essa abordagem acarreta três implicações práticas. A primeira é que concentra resolução na região da frequência fundamental e dos primeiros formantes, responsáveis pela identidade das vogais e por boa parte das pistas prosódicas e de sotaque. A segunda é que reduz o espectro de centenas de coeficientes por quadro para cerca de oitenta bandas, tornando tratável o mapeamento entre texto e áudio (Logan, 2000). A terceira é que, por ponderar o erro segundo a percepção, induz o modelo a acertar aquilo que o ouvinte de fato nota. Por isso o espectrograma Mel funciona como interface entre o modelo acústico, que prevê os quadros do espectrograma

Mel a partir do texto, e o *vocoder*, que os converte em forma de onda, permitindo que os dois sejam desenvolvidos de forma independente (Shen et al., 2018; Kong et al., 2020).

2.3.2 Representações Latentes e Tokens Acústicos

Modelos recentes também representam a fala por meio de representações latentes (codificações internas aprendidas pelas redes em espaços vetoriais reduzidos (Goodfellow et al., 2016)) e de tokens acústicos discretos, obtidos por quantização vetorial, como o *Vector Quantized Variational Autoencoder* (VQ-VAE) (van den Oord et al., 2017). Essa discretização permite tratar o áudio de forma análoga à linguagem natural, viabilizando arquiteturas inspiradas em modelos de linguagem. Modelos como XTTS e NaturalSpeech 3 adotam essas representações como componentes centrais (Casanova et al., 2024; Ju et al., 2024).

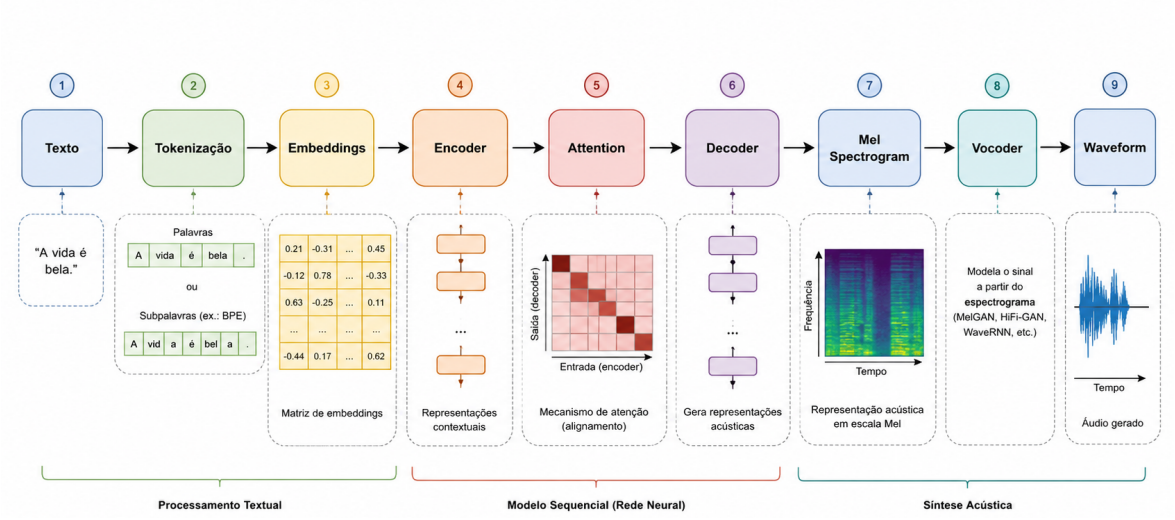
2.4 Componentes de Sistemas de Síntese de Fala com Aprendizagem Profunda

Um sistema de síntese de fala com aprendizagem profunda combina, de forma geral, os componentes descritos a seguir. A tokenização converte o texto de entrada em unidades discretas (caracteres, fonemas ou subpalavras) (Jurafsky e Martin, 2009), frequentemente apoiada pela conversão grafema-fonema (*Grapheme-to-Phoneme*, G2P), que explicita a pronúncia e é especialmente útil para sotaques regionais e palavras irregulares (Taylor, 2009). Cada token é mapeado em um *embedding*, vetor contínuo aprendido durante o treino; além dos *embeddings* textuais, sistemas modernos empregam *embeddings* de locutor, idioma ou sotaque para controlar características da fala (Vaswani et al., 2017; van den Oord et al., 2016).

O *encoder* transforma esses *embeddings* em representações linguísticas contextualizadas, e o *decoder* as converte em representações acústicas (tipicamente espectrogramas Mel), de forma autorregressiva ou paralela (Shen et al., 2018; Ren et al., 2019). Mecanismos de atenção alinham dinamicamente texto e fala, sendo a autoatenção dos *Transformers* eficiente na modelagem de dependências de longo alcance (Vaswani et al., 2017). Por fim, *vocoders* neurais (WaveNet, WaveGlow, HiFi-GAN, DiffWave) convertem as representações acústicas na forma de onda final, com qualidade perceptiva muito superior à dos *vocoders* paramétricos

clássicos (van den Oord et al., 2016; Prenger et al., 2019; Kong et al., 2020, 2021). A Figura 2.7 resume esse fluxo.

Figura 2.7: Visão geral do fluxo de um sistema de síntese de fala baseado em aprendizagem profunda, do texto à *waveform*.



Fonte: elaborado pelo autor (2026).

2.4.1 Checkpoints e Formato de Armazenamento

Em aprendizagem profunda, *checkpoints* são estados (pesos e metadados de treino) salvos durante o treinamento, que permitem retomar experimentos e, sobretudo, reutilizar modelos pré-treinados como ponto de partida para ajuste fino, reduzindo o custo computacional frente ao treinamento a partir de pesos aleatórios. Nesta dissertação, os *checkpoints* são armazenados no formato *safetensors*, amplamente adotado em plataformas como o Hugging Face[1] por evitar a execução de código durante o carregamento; esse formato também permitiu verificar diretamente as alterações de pesos entre as versões dos modelos ao longo dos experimentos.

2.5 Sotaques e Variação Linguística

Na tradição sociolinguística inaugurada por Labov (1972), o sotaque pode ser compreendido como o conjunto de características fonéticas, prosódicas e articulatórias que identificam a

[1] `https://huggingface.co`

origem regional, social ou cultural de um falante. Essas características manifestam-se principalmente na pronúncia das palavras, nos padrões de entonação, ritmo da fala e realização de determinados fonemas.

Na sociolinguística, sotaques são compreendidos como manifestações naturais da variação linguística presente nas línguas humanas. Como argumenta Bagno (2007), diferentes grupos sociais e regiões desenvolvem padrões específicos de fala influenciados por fatores históricos, culturais e geográficos. Dessa forma, sotaques não devem ser interpretados como desvios ou erros linguísticos, mas como formas legítimas de expressão linguística e identidade cultural.

No contexto computacional, sotaques representam um desafio importante para sistemas de processamento de fala, incluindo o reconhecimento automático e a síntese de fala, como evidencia o estudo de Batista (2019) sobre identificação automática de sotaques regionais brasileiros.

Em síntese de fala, a incorporação de sotaques torna-se particularmente relevante para promover maior naturalidade, identificação cultural e representatividade linguística em sistemas TTS (Tinchev et al., 2023; Zhou et al., 2023). Modelos capazes de reproduzir características regionais específicas podem contribuir para experiências mais inclusivas e adequadas ao contexto sociolinguístico dos usuários.

Na prática, contudo, sistemas TTS de uso corrente raramente oferecem vozes com sotaques não padronizados, sobretudo pela dificuldade de obter dados de treinamento de variedades pouco documentadas. Berger et al. (2025) mostram que esse obstáculo pode ser atenuado pelo compartilhamento de conhecimento entre variedades: um *frontend* neural treinado sobre múltiplos sotaques reduz em até 95% o volume de dados de pronúncia necessário para sotaques de poucos recursos.

2.5.1 Prosódia e Entonação

Segundo Cagliari (1992), a prosódia corresponde ao conjunto de características suprassegmentais da fala relacionadas à entonação, ritmo, duração e intensidade dos sons produzidos. Diferentemente dos fonemas individuais, a prosódia atua sobre sequências maiores da fala, influenciando diretamente a expressividade e interpretação do discurso.

A entonação representa um dos principais componentes prosódicos e refere-se às varia-

ções da frequência fundamental ao longo da fala. Essas variações contribuem para a transmissão de informações emocionais, pragmáticas e estruturais da linguagem, como discute Moraes (2008) para o caso do português brasileiro.

Em sistemas de síntese de fala, a modelagem prosódica exerce papel fundamental na percepção de naturalidade da fala sintetizada. Falhas relacionadas à prosódia frequentemente resultam em fala artificial, monótona ou pouco expressiva, mesmo quando a pronúncia fonética encontra-se correta.

Além disso, sotaques regionais frequentemente apresentam padrões prosódicos específicos, envolvendo alterações de ritmo, acentuação e contornos de entonação. Dessa forma, a representação adequada da prosódia constitui um elemento importante para a incorporação de regionalismos em sistemas *TTS*.

2.5.2 Variação Fonética

Conforme Ladefoged e Johnson (2012), a variação fonética refere-se às diferentes formas de realização acústica e articulatória dos sons da fala em uma língua. Essas variações podem ocorrer em função de fatores geográficos, sociais, históricos ou contextuais, influenciando diretamente a pronúncia dos fonemas.

No português brasileiro, diversos fenômenos fonéticos variam regionalmente, incluindo abertura vocálica, palatalização, redução vocálica e padrões específicos de realização consonantal, descritos por Silva (2011) em seu estudo sobre a fonética do idioma. Essas diferenças contribuem diretamente para a percepção de sotaques regionais distintos.

2.5.3 Regionalismos no Português Brasileiro

Como observa Castilho (2010), o português brasileiro apresenta elevada diversidade linguística devido à extensão territorial do país e à influência de diferentes processos históricos, culturais e migratórios. Como consequência, diferentes regiões desenvolveram padrões próprios de pronúncia, vocabulário, prosódia e construções linguísticas.

No contexto da fala, essas diferenças regionais manifestam-se principalmente por meio de sotaques específicos associados a diferentes estados e regiões do Brasil. Estudos sociolinguísticos apontam que tais variações não se limitam apenas à pronúncia isolada de fonemas,

envolvendo também padrões prosódicos, ritmo de fala e preferências lexicais (Bagno, 2007).

Apesar dessa diversidade, grande parte das bases de dados utilizadas em sistemas de síntese e reconhecimento de fala concentra-se em variedades linguísticas provenientes das regiões Sudeste e Sul do Brasil. Isso resulta em desbalanceamento regional nos conjuntos de treinamento e reduz a representatividade de sotaques menos presentes na literatura e na indústria (Batista, 2019).

Nesse contexto, a investigação de abordagens voltadas à modelagem de sotaques regionais específicos, como o sotaque paraibano, torna-se relevante tanto do ponto de vista tecnológico quanto sociolinguístico.

2.6 Ferramentas e Modelos Abertos de TTS

Modelos e ferramentas open source de síntese de fala têm desempenhado um papel importante na democratização do desenvolvimento de sistemas TTS, permitindo que pesquisadores e desenvolvedores realizem experimentos, adaptações e treinamentos sem dependência exclusiva de plataformas proprietárias. Frameworks abertos possibilitam acesso aos pesos dos modelos, arquiteturas neurais e pipelines de treinamento, favorecendo pesquisas relacionadas à adaptação linguística, sotaques regionais e personalização de fala.

Entre as principais ferramentas utilizadas na comunidade destacam-se projetos como Coqui TTS, Mozilla TTS, ESPnet-TTS e modelos recentes disponibilizados por meio da plataforma Hugging Face. Essas ferramentas frequentemente oferecem suporte para treinamento, ajuste fino e inferência de modelos neurais modernos de síntese de fala. Além de demonstrações e notebooks com códigos e receitas prontas para execução de ajustes finos destes modelos.

2.6.1 Pré-Treinamento

O pré-treinamento (*pre-training*) corresponde à etapa inicial de treinamento de modelos de síntese de fala utilizando grandes volumes de dados de fala. Durante essa fase, o sistema aprende padrões gerais relacionados à pronúncia, prosódia, alinhamento fonético e características acústicas da linguagem.

O processo parte de um grande corpus de pares texto-áudio e de um modelo cujos pesos

são inicializados de forma aleatória, isto é, sem qualquer conhecimento prévio sobre fala ou linguagem. O treinamento é um ciclo iterativo: o modelo tenta produzir o áudio correspondente a cada texto, a diferença entre a sua saída e o áudio real é medida por uma função de perda e os pesos são corrigidos gradualmente para reduzir essa diferença. Repetido milhões de vezes sobre dados variados, esse ciclo faz o modelo internalizar, em seus pesos, o conhecimento geral de como a língua soa (Goodfellow et al., 2016). O mesmo procedimento descreve o ajuste fino (Seção 2.6.2); a diferença essencial está no ponto de partida dos pesos, no corpus utilizado e na intensidade do treinamento.

Devido à elevada demanda computacional e à necessidade de grandes corpora, uma vez que todo o conhecimento do modelo precisa ser aprendido a partir do ruído inicial, o pré-treinamento geralmente é realizado pelas empresas, universidades ou grupos de pesquisa responsáveis pelo desenvolvimento dos modelos-base. Após essa etapa, os pesos resultantes são salvos em *checkpoints* (Seção 2.4.1) e podem ser disponibilizados publicamente para reutilização em tarefas de adaptação e personalização.

2.6.2 Ajuste Fino

O ajuste fino (*fine-tuning*) consiste no processo de adaptação de um modelo previamente treinado para um domínio específico utilizando conjuntos menores de dados. Em sistemas de síntese de fala, essa abordagem é frequentemente utilizada para adaptar modelos para novos idiomas, estilos de fala, locutores específicos ou sotaques regionais.

Na prática, o ajuste fino repete o mesmo ciclo de treinamento do pré-treinamento (Seção 2.6.1), com três diferenças. Primeiro, os pesos não partem de valores aleatórios, mas do checkpoint pré-treinado: o modelo já sabe falar, e o objetivo é apenas refinar esse conhecimento. Segundo, o corpus de treinamento é substituído por dados do domínio alvo, tipicamente muito menor que o corpus de pré-treino. Terceiro, o treinamento é mais curto e cauteloso (taxas de aprendizado menores e menos épocas), de modo que os pesos se desloquem apenas o suficiente para incorporar o novo comportamento sem apagar o que o modelo já aprendeu. Diferentemente da clonagem *zero-shot* (Seção 2.6.3), aqui os pesos são efetivamente modificados: a comparação entre os *checkpoints* antes e depois do processo revela alterações mensuráveis, ainda que pequenas.

Esse equilíbrio delicado explica um risco conhecido do ajuste fino: o esquecimento catas-

trófico (*catastrophic forgetting*) (Kirkpatrick et al., 2017), no qual o treino sobre um conjunto pequeno e pouco diverso degrada capacidades gerais aprendidas no pré-treino, como a pronúncia de palavras ausentes dos novos dados ou a estabilidade da síntese em textos longos. Por outro lado, um treinamento conservador demais pode produzir o efeito oposto: os pesos mudam, mas o comportamento do modelo não se altera de forma audível.

Quanto à extensão da atualização, distinguem-se três variantes. No ajuste fino completo, todos os pesos do modelo são atualizados. No ajuste fino parcial, componentes do modelo são congelados (por exemplo, o codificador de texto ou o vocoder), e apenas os módulos restantes recebem atualizações, reduzindo custo de memória e risco de esquecimento. Em métodos eficientes em parâmetros (do inglês, *Parameter-Efficient Fine-Tuning*, PEFT), como o LoRA (*Low-Rank Adaptation*), todos os pesos originais permanecem congelados e a adaptação é aprendida em pequenas matrizes adicionais acopladas ao modelo, de modo que apenas uma fração mínima dos parâmetros totais é treinada (Hu et al., 2022).

Por aproveitar os padrões gerais já aprendidos no pré-treinamento com menor demanda de dados e de computação, o ajuste fino é particularmente relevante em cenários de escassez de dados, como o de sotaques regionais menos representados, e possui papel central nesta pesquisa.

2.6.3 Clonagem de Voz

A clonagem de voz (*voice cloning*) consiste na geração de novos enunciados com características vocais semelhantes às de um locutor de referência (Tan et al., 2021); em sistemas neurais modernos, poucos segundos de gravação podem ser suficientes, o que popularizou a abordagem em aplicações de personalização de fala e síntese expressiva.

Na modalidade *zero-shot*, investigada nesta dissertação, os pesos do modelo permanecem congelados durante a inferência. O áudio de referência é processado por um codificador de voz ou codificador acústico, responsável por extrair uma representação vetorial da identidade vocal, como um *embedding* de locutor ou uma sequência de características acústicas (Casanova et al., 2024; Chen et al., 2025). Essa representação contém informações relacionadas ao timbre, à faixa de frequência, à entonação, ao ritmo e a outros aspectos da voz.

Em seguida, a representação da voz é utilizada para condicionar o modelo TTS juntamente com o texto de entrada. Dessa forma, o modelo emprega o conhecimento geral sobre

produção de fala, adquirido durante o pré-treinamento, para gerar uma nova representação acústica com o conteúdo textual desejado e com características semelhantes às da voz de referência; por fim, um vocoder converte essa representação em um sinal de áudio (Tan et al., 2021).

Nesse processo, a voz do locutor não é incorporada permanentemente aos pesos do modelo, pois a representação vocal é produzida apenas a partir do áudio fornecido na inferência; essa propriedade pode ser verificada empiricamente comparando os *checkpoints* antes e depois da geração (Seção 2.4.1), que permanecem idênticos. No ajuste fino, por outro lado, o modelo é treinado com várias amostras de uma determinada voz, permitindo que parte de suas características seja incorporada aos pesos atualizados ou a adaptadores específicos (Tokuda et al., 2013; Hu et al., 2022). Assim, a principal diferença entre a clonagem por referência e o ajuste fino está no fato de que, na primeira, a voz atua como condicionamento temporário, enquanto, na segunda, suas características são aprendidas durante o treinamento (Seção 2.6.2).

Uma consequência direta dessa formulação é relevante para esta dissertação: como as pistas extraídas da referência capturam características globais e entrelaçadas da voz, não há garantia de que traços específicos de sotaque, sobretudo os fonéticos e prosódicos discutidos na Seção 2.5, sejam transferidos junto com o timbre. A capacidade de reproduzir sotaque depende de tais traços já estarem representados no conhecimento adquirido durante o pré-treinamento, o que por sua vez depende da diversidade dos dados vistos nessa etapa. Essa limitação motiva os experimentos de clonagem descritos na Seção 4.5.1.

2.7 Avaliação de Sistemas de Síntese de Fala

2.7.1 Avaliação Perceptual

A avaliação da qualidade da fala sintetizada constitui um componente central no desenvolvimento, comparação e validação de sistemas de síntese de fala. A qualidade perceptiva da fala envolve fatores subjetivos associados à experiência humana, como naturalidade, inteligibilidade, fluidez prosódica e ausência de artefatos audíveis. Por esse motivo, a avaliação em síntese de fala tradicionalmente combina métodos subjetivos, baseados no julgamento

humano, e métodos automáticos, que buscam estimar a qualidade de forma reprodutível e escalável.

2.7.2 *Mean Opinion Score* (MOS)

Entre os métodos subjetivos, destaca-se o *Mean Opinion Score* (MOS), padronizado pela recomendação ITU-T P.800 (ITU-T, 1996). O MOS permite capturar a percepção global da qualidade da fala sintetizada a partir da opinião de ouvintes humanos, sendo amplamente utilizado para comparar diferentes sistemas, arquiteturas e configurações experimentais.

De forma geral, a avaliação por MOS é realizada por meio da apresentação de amostras de áudio sintetizado a um conjunto de ouvintes humanos, que atribuem notas discretas conforme sua percepção subjetiva da qualidade da fala. As avaliações são feitas em uma escala ordinal de cinco pontos, na qual valores mais baixos indicam baixa qualidade perceptiva e valores mais altos representam fala considerada natural, clara e agradável. O valor final de MOS corresponde à média aritmética das notas atribuídas pelos participantes para cada amostra avaliada, podendo ser acompanhado por medidas estatísticas de dispersão, como desvio padrão ou intervalos de confiança.

Tabela 2.1: Exemplo de escala de *Mean Opinion Score* (MOS) para avaliação de inteligibilidade da fala sintetizada.

Valor	**Descrição da avaliação**
1	Muito ruim (fala ininteligível)
2	Ruim (fala difícil de compreender)
3	Regular (inteligibilidade aceitável, com esforço)
4	Boa (fala clara, poucos problemas)
5	Excelente (fala totalmente clara e inteligível)

Pergunta ao avaliador: "Quão inteligível é a fala apresentada neste áudio?"

Fonte: elaborada pelo autor (2026), com base na escala de ITU-T (1996).

Instrumentos do tipo *survey*, como o MOS, são amplamente empregados na avaliação perceptiva de sistemas de síntese de fala, capturando aspectos globais como inteligibilidade, naturalidade e qualidade. Apesar de sua robustez, dependem da participação de ouvintes humanos, o que motivou o desenvolvimento de alternativas automáticas, discutidas a seguir.

2.7.3 Preditores Automáticos de MOS

Com o objetivo de reduzir o custo e a dependência de avaliações humanas em fases iniciais de desenvolvimento, têm sido propostos modelos automáticos de predição de MOS, frequentemente denominados *MOS predictors* (preditores de MOS) (Lo et al., 2019). Esses modelos utilizam técnicas de aprendizado profundo para estimar a qualidade perceptiva da fala sintetizada diretamente a partir do sinal de áudio, sem a necessidade de referências externas ou participação humana direta.

Embora não substituam avaliações subjetivas formais, os preditores de MOS são amplamente empregados como ferramenta de avaliação preliminar, permitindo comparações rápidas entre diferentes modelos, configurações e estágios de treinamento. Dessa forma, esses métodos possibilitam uma triagem inicial mais eficiente, reservando avaliações humanas detalhadas para experimentos finais ou comparativos, nos quais a validação perceptiva rigorosa é essencial (Zezario et al., 2023).

Entre os preditores disponíveis na literatura, esta dissertação adota dois de naturezas complementares. O *NISQA*, proposto por Mittag et al. (2021), é um modelo baseado em redes convolucionais com mecanismo de autoatenção que, além de estimar a qualidade global da fala, decompõe a avaliação em dimensões interpretáveis, como ruído percebido, descontinuidade, coloração e intensidade sonora. O *DNSMOS*, desenvolvido por Reddy et al. (2022), por sua vez, é um preditor não intrusivo proposto originalmente para avaliar supressores de ruído, que fornece estimativas separadas para a qualidade do sinal de voz, do ruído de fundo e da qualidade global, seguindo a recomendação ITU-T P.835. A utilização conjunta dos dois preditores permite uma avaliação automática mais robusta, combinando uma medida global com dimensões específicas da qualidade perceptiva.

2.7.4 Avaliação Baseada em STT e Distância de Levenshtein

Além do MOS, métricas auxiliares podem ser empregadas para analisar aspectos específicos da síntese de fala, em particular a preservação do conteúdo linguístico gerado pelo sistema. Uma dessas abordagens consiste na utilização de transcrição automática da fala sintetizada, seguida do cálculo da distância de Levenshtein (Levenshtein, 1966) entre o texto original de entrada e o texto reconhecido por um sistema de reconhecimento automático de fala (STT).

Durante o desenvolvimento desta pesquisa, observou-se que esse tipo de métrica poderia ser explorado como um indicador indireto de inteligibilidade, motivado pela necessidade de comparações automáticas simples e reprodutíveis entre diferentes configurações de síntese, especialmente em estágios iniciais de experimentação.

O princípio dessa métrica baseia-se no fluxo em que o texto original é convertido em fala por um sistema TTS, o áudio é transcrito automaticamente e a similaridade entre o texto de entrada e o texto reconhecido é quantificada por meio da distância de Levenshtein, normalizada para um valor contínuo. Valores próximos de 1 indicam alta correspondência textual; valores menores sugerem perda de inteligibilidade ou alteração do conteúdo linguístico. Essa abordagem não captura prosódia nem sotaque, mas permite comparações automáticas escaláveis entre sistemas.

Formalmente, a distância de Levenshtein entre duas cadeias de caracteres $a = a_1 a_2 \ldots a_n$ e $b = b_1 b_2 \ldots b_m$ é definida como o número mínimo de operações de edição (inserções, remoções e substituições de um único caractere) necessárias para transformar a em b. O cálculo é feito por programação dinâmica: preenche-se uma matriz em que cada célula registra a distância entre um prefixo de a e um prefixo de b, obtida a partir das células vizinhas somando o custo de uma remoção, de uma inserção ou de uma substituição; a célula do canto inferior direito acumula, ao final, a distância entre as cadeias completas (Levenshtein, 1966; Wagner e Fischer, 1974). A Figura 2.8 ilustra esse preenchimento para as cadeias “mainha” e “minha”: as células sombreadas marcam o caminho de custo mínimo e a última célula contém a distância final, $\mathrm{lev}(a, b) = 1$, correspondente a uma única remoção.

Figura 2.8: Matriz de programação dinâmica da distância de Levenshtein entre a referência "mainha" e a transcrição "minha".

referência (a = "mainha") \ *transcrição (b = "minha")*	ε	**m**	**i**	**n**	**h**	**a**
ε	0	1	2	3	4	5
m	1	0	1	2	3	4
a	2	1	1	2	3	3
i	3	2	1	2	3	4
n	4	3	2	1	2	3
h	5	4	3	2	1	2
a	6	5	4	3	2	1

Fonte: elaborado pelo autor (2026).

Como a distância absoluta cresce com o comprimento das cadeias, o que dificulta a comparação entre sentenças de tamanhos distintos, adota-se nesta dissertação a versão normalizada (Marzal e Vidal, 1993), convertida em uma medida de similaridade no intervalo $[0, 1]$, apresentada na Equação 2.1:

$$\mathrm{sim}(a, b) = 1 - \frac{\mathrm{lev}(a, b)}{\max(n,\ m)}, \tag{2.1}$$

de modo que $\mathrm{sim}(a, b) = 1$ indica transcrição idêntica ao texto de entrada e valores progressivamente menores indicam maior divergência lexical. No exemplo da Figura 2.8, a similaridade é $\mathrm{sim} = 1 - 1/6 \approx 0{,}83$.

A aplicação concreta dessas métricas aos modelos adaptados nesta dissertação (amostras de teste, Whisper, preditores de MOS, procedimento e métricas de treino) está descrita na Seção 4.6.

2.7.5 Testes de Significância Estatística

Tanto as métricas automáticas quanto as respostas de ouvintes humanos são medidas sujeitas a variabilidade amostral: uma diferença observada entre duas condições pode refletir um efeito real da adaptação ou apenas flutuação aleatória. Para distinguir os dois casos, esta

dissertação emprega testes de hipóteses. Em cada teste, formula-se uma hipótese nula H_0 que representa a ausência de efeito (por exemplo, "as duas condições são equivalentes") e calcula-se o p-valor, definido como a probabilidade de observar um resultado tão ou mais extremo que o obtido, supondo H_0 verdadeira. Quando o p-valor é inferior ao nível de significância adotado ($\alpha = 0{,}05$ em todas as análises), rejeita-se H_0 e considera-se a diferença estatisticamente significativa (Agresti, 2013). Três tipos de testes são utilizados nesta pesquisa, os quais são escolhidos conforme a natureza dos dados. Os testes são discutidos a seguir.

O teste binomial exato é aplicado quando cada observação é binária. Esse é o caso da avaliação perceptiva par a par (Seção 4.7): descartadas as respostas de empate e de indistinguibilidade, cada resposta restante equivale a um ensaio de Bernoulli com dois desfechos (favorável ao áudio adaptado ou ao original). Sob a hipótese nula de que os dois áudios são indistinguíveis quanto ao critério avaliado, cada desfecho tem probabilidade $p = 0{,}5$, e o número total de respostas favoráveis em n ensaios segue uma distribuição Binomial$(n,\ 0{,}5)$. O teste exato calcula o p-valor diretamente a partir dessa distribuição, somando as probabilidades de todos os desfechos tão ou mais extremos que o observado, sem recorrer a aproximações assintóticas, o que o torna adequado inclusive para amostras pequenas (Agresti, 2013).

Para comparar proporções entre grupos independentes (por exemplo, verificar se a taxa de respostas favoráveis de um modelo é maior que a de outro), utiliza-se o teste exato de Fisher, que avalia a associação em uma tabela de contingência 2×2 calculando a probabilidade exata de cada configuração da tabela sob a hipótese nula de proporções iguais (Agresti, 2013). Quando várias comparações são realizadas simultaneamente, a probabilidade de algum falso positivo aumenta; nesses casos, os p-valores são ajustados pela correção de Holm, procedimento sequencial que controla a taxa de erro global da família de testes e é uniformemente mais poderoso que a correção de Bonferroni clássica (Holm, 1979).

Para as métricas automáticas contínuas (similaridade de Levenshtein e escores dos preditores de MOS), as condições original e ajustada são comparadas pelo teste de postos sinalizados de Wilcoxon (*Wilcoxon signed-rank test*) (Wilcoxon, 1945), na versão pareada: cada sentença de teste gera um par de escores (original e ajustado), e o teste avalia se a mediana das diferenças é nula. Trata-se de um teste não paramétrico, baseado nos postos das dife-

renças, que não pressupõe normalidade das distribuições, suposição pouco plausível para escores de preditores de MOS, que são limitados ao intervalo da escala e frequentemente assimétricos. O pareamento por sentença isola o efeito da condição, eliminando a variabilidade devida ao conteúdo textual.

2.8 Síntese do Capítulo

Este capítulo reuniu os fundamentos necessários à compreensão da pesquisa. Percorreu-se a evolução da síntese de fala, dos sistemas concatenativos e paramétricos às arquiteturas de redes neurais atuais. Em seguida, foram apresentados os conceitos de redes neurais e de *Transformers*, bem como as representações de áudio empregadas por esses modelos, com destaque para o espectrograma Mel. Foram então descritos os componentes de um sistema de síntese e as estratégias de reaproveitamento de modelos pré-treinados, a saber, o ajuste fino e a clonagem de voz, além da base sociolinguística relativa a sotaques e regionalismos do português brasileiro e das abordagens de avaliação, subjetivas, automáticas e estatísticas, adotadas nesta dissertação. No próximo capítulo é apresentada a revisão da literatura, que situa esta pesquisa em relação aos trabalhos que aplicaram esses fundamentos à adaptação de modelos e à representação de sotaques.

Capítulo 3

Revisão da Literatura

Este capítulo apresenta a revisão da literatura relacionada à síntese de fala com aprendizagem de máquina, com foco em arquiteturas modernas de TTS, modelagem de sotaques e recursos disponíveis para o português brasileiro. A revisão tem como objetivo contextualizar os avanços recentes da área, identificar abordagens utilizadas na incorporação de variações regionais em sistemas de síntese de fala e analisar limitações ainda existentes na representação de sotaques do português brasileiro.

Inicialmente, são apresentados os fundamentos da síntese de fala e a evolução das principais abordagens utilizadas ao longo do desenvolvimento da área. Em seguida, são discutidas arquiteturas neurais contemporâneas aplicadas a sistemas TTS, incluindo modelos autorregressivos, não autorregressivos e baseados em difusão.

Posteriormente, são analisados trabalhos relacionados à síntese de fala para o português brasileiro, com ênfase em modelos multilíngues e estratégias de adaptação para o idioma. Também são discutidos estudos voltados à modelagem de sotaques e variações regionais, abordando aspectos fonéticos, prosódicos e sociolinguísticos relevantes para a síntese de fala.

Por fim, são apresentados corpora de fala disponíveis para o português brasileiro, especialmente aqueles relacionados à diversidade regional da fala, bem como as limitações identificadas na literatura quanto à representação de sotaques específicos, como o sotaque paraibano.

3.1 Metodologia da Revisão Bibliográfica

Esta revisão bibliográfica teve como objetivo identificar arquiteturas, estratégias e bases de dados relacionadas à síntese de fala baseada em aprendizagem profunda, com foco em sotaques regionais e português brasileiro. A revisão buscou reunir trabalhos relevantes sobre modelos modernos de *Text-to-Speech* (TTS), adaptação de sotaque, clonagem de voz e corpora de fala aplicados à síntese de fala.

3.1.1 Questões Norteadoras

A revisão bibliográfica foi conduzida a partir das seguintes questões norteadoras:

- Quais arquiteturas recentes têm sido utilizadas em sistemas de síntese de fala baseados em aprendizagem profunda?
- Quais bases de dados em português brasileiro estão disponíveis para treinamento e avaliação de modelos de síntese de fala?
- Existem trabalhos voltados especificamente à representação de sotaques regionais brasileiros em sistemas TTS?

3.1.2 Estratégia de Busca

As buscas foram realizadas em bases e repositórios científicos amplamente utilizados na área de computação, incluindo Google Scholar, IEEE Xplore, ACM Digital Library, arXiv e Hugging Face. Também foram consultados repositórios públicos e documentações técnicas relacionadas a modelos abertos de síntese de fala.

Foram utilizadas combinações de palavras-chave em inglês e português, incluindo os termos “Text-to-Speech”, “Deep Learning Speech Synthesis”, “Speech Synthesis”, “Accent Transfer”, “Voice Cloning”, “Zero-shot TTS”, “Brazilian Portuguese TTS”, “Regional Accent” e “Speech Corpus”.

3.1.3 Critérios de Inclusão

Foram incluídos trabalhos que atendessem aos seguintes critérios:

- Trabalhos relacionados à síntese de fala baseada em aprendizagem profunda;
- Estudos envolvendo sotaques, adaptação de locutor ou transferência de sotaque;
- Trabalhos relacionados ao português brasileiro ou síntese multilíngue;
- Artigos publicados em periódicos, conferências ou repositórios científicos reconhecidos;
- Estudos contendo descrição metodológica suficiente para análise e reprodução, preferencialmente com código aberto.

3.1.4 Critérios de Exclusão

Foram excluídos:

- Trabalhos sem relação direta com síntese de fala ou processamento de fala;
- Artigos sem descrição técnica ou metodológica adequada;
- Trabalhos cujo texto completo não estava disponível para consulta;
- Modelos candidatos aos experimentos de adaptação desta dissertação que não disponibilizassem código e pesos abertos para reprodução;
- Estudos publicados antes de 2016, exceto trabalhos clássicos relevantes para a contextualização histórica da área e exceto corpora que possam eventualmente ser utilizados no contexto desta dissertação.

3.1.5 Processo de Seleção dos Estudos

Inicialmente, foi realizada uma busca exploratória utilizando as palavras-chave definidas previamente. Em seguida, os trabalhos identificados passaram por uma etapa de triagem baseada na leitura de títulos, resumos e palavras-chave, visando verificar sua relevância para os objetivos da pesquisa.

Posteriormente, os estudos selecionados foram analisados de forma mais detalhada, considerando aspectos como arquitetura utilizada, idioma suportado, presença de modelagem

de sotaque, disponibilidade de código ou modelos abertos e utilização de bases de dados em português brasileiro.

A partir dessa análise, foram selecionados os trabalhos considerados mais relevantes para fundamentar a discussão sobre síntese de fala baseada em aprendizagem profunda, adaptação de sotaques e representação de variações regionais em sistemas TTS.

3.2 Modelos de Síntese de Fala Baseados em Aprendizagem Profunda

Os avanços recentes em aprendizagem profunda transformaram significativamente a área de síntese de fala, permitindo a construção de sistemas capazes de gerar fala com maior naturalidade, inteligibilidade e expressividade. Diferentemente das abordagens concatenativas e paramétricas tradicionais, os modelos modernos passaram a aprender diretamente representações acústicas complexas a partir de grandes volumes de dados, reduzindo a necessidade de regras linguísticas explícitas e engenharia manual de características.

O desenvolvimento dessas arquiteturas possibilitou avanços importantes na modelagem temporal da fala, alinhamento texto-áudio, controle prosódico e adaptação a diferentes locutores e idiomas. Além disso, novas abordagens vêm permitindo síntese *zero-shot*, clonagem de voz e incorporação de sotaques regionais utilizando quantidades reduzidas de dados.

Nesta seção, são apresentados alguns dos principais modelos e arquiteturas que contribuíram para a evolução recente da síntese de fala baseada em aprendizagem profunda.

3.2.1 WaveNet

Um dos marcos mais importantes da síntese de fala moderna foi o modelo WaveNet, proposto por van den Oord et al. (2016). O modelo introduziu uma abordagem autorregressiva[1] capaz de gerar diretamente a forma de onda do áudio, amostra por amostra, utilizando redes neurais convolucionais profundas com convoluções dilatadas.

Diferentemente das abordagens paramétricas tradicionais, o WaveNet modela direta-

[1] Os modelos autorregressivos produzem saídas passo a passo, usando valores anteriores, enquanto os não autorregressivos geram toda a sequência de uma só vez.

mente a distribuição probabilística do sinal de áudio, permitindo capturar características acústicas complexas da fala humana. Essa formulação possibilitou um salto significativo na naturalidade perceptiva da síntese de fala, aproximando os resultados produzidos da fala humana natural.

A arquitetura utiliza convoluções causais dilatadas para ampliar o campo receptivo da rede sem aumento excessivo do custo computacional, permitindo modelar dependências temporais de longo alcance no sinal de fala. Além disso, o modelo pode ser condicionado por informações linguísticas e acústicas, como espectrogramas mel ou *embeddings* de locutor.

Apesar da elevada qualidade obtida, o WaveNet apresenta alto custo computacional durante a inferência devido ao seu caráter autorregressivo, o que motivou o desenvolvimento posterior de arquiteturas mais eficientes para síntese de fala em tempo real.

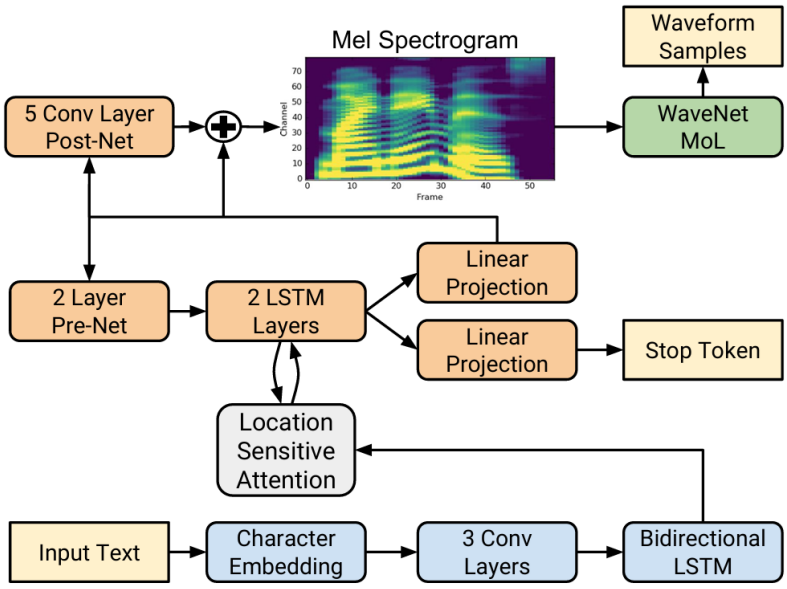
3.2.2 Tacotron e Tacotron 2

Posteriormente, arquiteturas ponta a ponta passaram a integrar diretamente o processo de conversão texto-fala. Entre os modelos mais influentes dessa abordagem destaca-se o Tacotron, posteriormente evoluído para o Tacotron 2 por Shen et al. (2018).

O Tacotron 2 combina redes neurais recorrentes e mecanismos de atenção para converter sequências textuais diretamente em espectrogramas mel, eliminando grande parte das etapas intermediárias presentes em pipelines tradicionais de síntese de fala.

A arquitetura, ilustrada na Figura 3.1, é composta por um codificador textual responsável por extrair representações linguísticas da sequência de entrada, um mecanismo de atenção que realiza o alinhamento entre texto e fala e um decodificador autorregressivo que gera os quadros acústicos do espectrograma mel. Posteriormente, um *vocoder* neural é utilizado para converter o espectrograma em forma de onda.

Figura 3.1: Arquitetura do Tacotron, ilustrando o fluxo do texto de entrada para a predição do espectrograma mel e posterior geração de *waveform* via *vocoder*.



Fonte: Shen et al. (2018).

A utilização de mecanismos de atenção permitiu que o modelo aprendesse automaticamente o alinhamento entre texto e áudio sem necessidade de alinhamentos fonéticos explícitos. Essa característica tornou o Tacotron uma das arquiteturas mais influentes da síntese de fala moderna, servindo de base para diversos trabalhos posteriores.

Entretanto, devido ao seu caráter autorregressivo, o Tacotron ainda apresenta limitações relacionadas à velocidade de inferência e estabilidade de alinhamento em sequências longas.

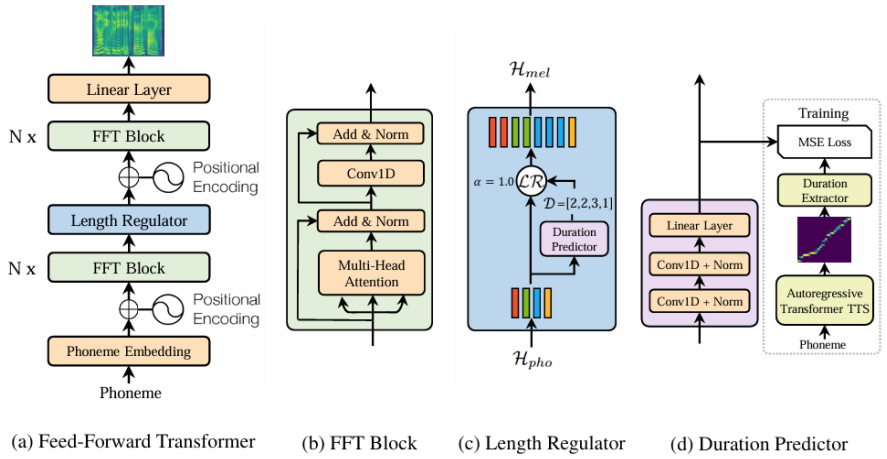
3.2.3 FastSpeech

Com o objetivo de reduzir a latência de inferência e aumentar a estabilidade da síntese de fala, surgiram modelos não autorregressivos, entre os quais se destaca o FastSpeech, proposto por Ren et al. (2019).

Diferentemente de arquiteturas autorregressivas, o FastSpeech realiza a geração paralela da sequência acústica, eliminando a dependência temporal entre os quadros de saída. Para isso, o modelo substitui o mecanismo de atenção dinâmica por um regulador de comprimento

(*length regulator*), responsável por expandir as representações fonéticas de acordo com as durações previstas para cada unidade linguística.

Figura 3.2: Arquitetura geral do FastSpeech: (a) *Transformer feed-forward*; (b) bloco FFT; (c) regulador de comprimento; (d) preditor de duração.



(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

Fonte: Ren et al. (2019).

A arquitetura do FastSpeech (Figura 3.2) utiliza blocos *Transformer feed-forward* (blocos FFT, com atenção multi-cabeça e convoluções) para extração das representações linguísticas e um preditor de duração supervisionado, otimizado por perda MSE entre durações previstas e extraídas, para modelar explicitamente o alinhamento temporal entre texto e fala. Essa abordagem reduz significativamente o tempo de inferência e melhora a estabilidade do treinamento quando comparada a modelos autorregressivos.

Além disso, variantes posteriores, como o FastSpeech 2 (Ren et al., 2021), incorporaram mecanismos adicionais para modelagem explícita de *pitch*, energia e duração, ampliando o controle prosódico da síntese.

3.2.4 Modelos Baseados em Difusão

Mais recentemente, modelos baseados em processos de difusão passaram a ser explorados na síntese de fala. Essas arquiteturas utilizam processos iterativos de remoção de ruído para reconstrução progressiva das representações acústicas, permitindo elevada qualidade perceptiva e maior estabilidade de treinamento.

Entre os trabalhos representativos dessa abordagem destacam-se o Grad-TTS, de Popov et al. (2021), e o DiffWave, de Kong et al. (2021). O Grad-TTS utiliza modelos de difusão para geração de espectrogramas mel condicionados pelo texto, enquanto o DiffWave atua como vocoder neural baseado em difusão para reconstrução da forma de onda.

Os modelos baseados em difusão apresentam vantagens relacionadas à qualidade da síntese, modelagem detalhada das características acústicas e maior robustez durante o treinamento. Entretanto, em muitos casos, o processo iterativo de difusão ainda implica maior custo computacional durante a inferência quando comparado a arquiteturas puramente não autorregressivas.

Apesar disso, essas abordagens vêm se consolidando como uma das principais linhas de pesquisa recentes em síntese de fala baseada em aprendizagem profunda.

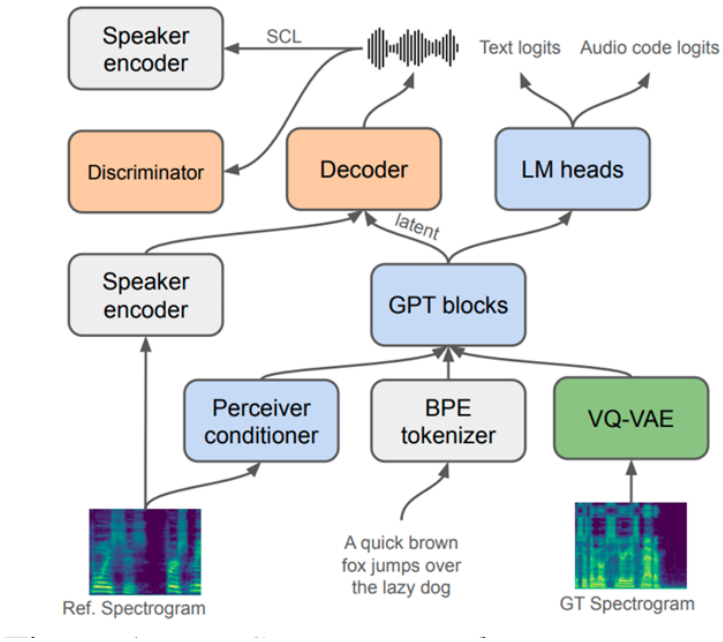
3.2.5 Modelos Recentes de Grande Escala

O avanço de modelos de grande escala impulsionou o desenvolvimento de sistemas de síntese de fala capazes de operar em cenários multilíngues, *zero-shot* e com clonagem de voz.

Entre os exemplos mais relevantes encontra-se o NaturalSpeech 3, apresentado por Ju et al. (2024), que combina aprendizado auto-supervisionado, modelagem hierárquica e mecanismos de difusão para separação explícita de atributos da fala, como conteúdo linguístico, prosódia e identidade vocal.

Além disso, modelos abertos recentes como o XTTS, de Casanova et al. (2024), o F5-TTS, de Chen et al. (2025), e o Fish-Speech, de Liao et al. (2024), passaram a disponibilizar suporte multilíngue e mecanismos de adaptação de locutor utilizando pequenas amostras de áudio de referência. No caso do XTTS, a arquitetura de treinamento (Figura 3.3) é composta por um codificador neural baseado em VQ-VAE para discretização acústica, um *Transformer* do tipo GPT-2 condicionado por *embeddings* extraídos de espectrogramas de referência e um decodificador baseado em *vocoder* neural para reconstrução da forma de onda.

Figura 3.3: Visão geral da arquitetura de treinamento do XTTS.



Fonte: Casanova et al. (2024).

Esses modelos vêm sendo explorados em tarefas de clonagem de voz, adaptação de sotaques e síntese multilíngue, permitindo maior flexibilidade e redução da quantidade de dados necessária para adaptação a novos locutores ou idiomas.

Além disso, trabalhos recentes investigam arquiteturas híbridas envolvendo *neural codecs*, modelos latentes e estratégias pseudo-autorregressivas, buscando equilibrar qualidade perceptiva, velocidade de inferência e controle da síntese (Yang et al., 2025).

O crescimento contínuo dessa área evidencia que a síntese de fala baseada em aprendizagem profunda permanece em rápida evolução, com o surgimento frequente de novas arquiteturas e estratégias de modelagem.

3.2.6 Orpheus TTS e Unsloth

O Orpheus TTS, desenvolvido pela Canopy Labs (2025), é um modelo recente de síntese de fala voltado à clonagem de voz e síntese multilíngue utilizando arquiteturas modernas baseadas em modelos generativos. O sistema busca produzir fala com elevada naturalidade utilizando pequenas amostras de áudio de referência, permitindo adaptação eficiente de locutor e geração zero-shot.

Além disso, o framework Unsloth, de Han et al. (2023), vem sendo utilizado em conjunto com modelos TTS para viabilizar estratégias de ajuste fino com baixo custo computacional. A ferramenta implementa otimizações de memória e treinamento utilizando técnicas como o LoRA (*Low-Rank Adaptation*), proposto por Hu et al. (2022), permitindo que apenas pequenas parcelas dos parâmetros do modelo sejam ajustadas durante o treinamento. Essa abordagem reduz significativamente o consumo de memória, tempo de treinamento e custo computacional quando comparada ao ajuste fino completo do modelo.

Dessa forma, modelos compatíveis com LoRA tornam-se candidatos relevantes para experimentos de adaptação de sotaque e síntese regionalizada, especialmente em cenários com infraestrutura computacional limitada e conjuntos de dados reduzidos.

3.2.7 Qwen3-TTS

O Qwen3-TTS, apresentado por Hu et al. (2026), representa uma geração recente de modelos TTS de grande escala voltados à síntese multilíngue e expressiva. O modelo combina capacidades multimodais e arquiteturas modernas de modelagem de áudio para permitir síntese de fala com elevada naturalidade, controle prosódico e clonagem de voz.

Além do suporte multilíngue, o modelo busca incorporar mecanismos de controle de estilo e expressividade, permitindo aplicações relacionadas à síntese emocional, adaptação de locutor e geração de fala em diferentes contextos comunicativos. Essas características tornam modelos dessa categoria relevantes para pesquisas envolvendo sotaques regionais e adaptação linguística em sistemas TTS.

3.3 Bases de Dados de Fala para o Português Brasileiro

A disponibilidade de bases de dados de fala constitui um dos principais fatores para o desenvolvimento e avaliação de sistemas modernos de síntese de fala baseados em aprendizagem profunda. Em tarefas de TTS, a qualidade, diversidade e quantidade dos dados utilizados influenciam diretamente aspectos como naturalidade, inteligibilidade, prosódia e capacidade de generalização dos modelos.

No contexto do português brasileiro, diferentes iniciativas vêm buscando disponibilizar corpora de fala voltados tanto para tarefas gerais de processamento de voz quanto para aplica-

ções específicas envolvendo síntese de fala, reconhecimento automático de fala e modelagem de sotaques regionais.

3.3.1 Corpora Públicos

Diversos corpora públicos vêm sendo utilizados em pesquisas envolvendo processamento de fala para o português brasileiro. Esses conjuntos de dados diferem em aspectos como quantidade de locutores, duração total de áudio, qualidade das gravações, estilo de fala e cobertura linguística.

Entre os recursos disponíveis, destaca-se o Corpus Linguístico da Paraíba (CoLingPB), organizado por Stein et al. (2013) na Universidade Federal da Paraíba, composto por entrevistas e registros de fala espontânea produzidos por falantes da região Nordeste do Brasil. O corpus apresenta relevância para estudos fonéticos, prosódicos e sociolinguísticos, especialmente em investigações relacionadas à variação regional da fala.

Outra iniciativa relevante é o projeto Sotaque Brasileiro, de Milan (2021), desenvolvido com o objetivo de registrar a diversidade regional da fala no Brasil por meio de coleta colaborativa de gravações. O corpus foi construído a partir da leitura de frases padronizadas por falantes de diferentes regiões do país, permitindo maior controle do conteúdo linguístico e comparabilidade entre amostras.

Além disso, trabalhos recentes também vêm explorando grandes bases derivadas automaticamente da *web*. Um exemplo é o *dataset* TAGARELA, apresentado por Oliveira et al. (2026), composto por milhares de horas de áudio extraídas de *podcasts* em português. Esse tipo de abordagem busca ampliar a escala dos corpora disponíveis para treinamento de modelos de aprendizagem profunda.

Modelos recentes de síntese de fala multilíngue também passaram a utilizar grandes volumes de dados em português brasileiro durante o pré-treinamento. O XTTS (Casanova et al., 2024), por exemplo, disponibiliza suporte ao português brasileiro utilizando bases multilíngues compostas por milhares de horas de fala provenientes de diferentes idiomas e locutores (Tabela 3.1).

Tabela 3.1: Número de horas de áudio por idioma no conjunto de dados utilizado no treinamento do XTTS.

Idioma	**Horas**	**Idioma**	**Horas**
Inglês	14.513,1	Tcheco	52,4
Alemão	3.584,4	Coreano	539,1
Espanhol	1.514,3	Húngaro	62,0
Francês	2.215,5	Japonês	57,3
Italiano	1.296,6	Turco	165,3
Português	2.386,8	Árabe	240,9
Russo	147,1	Chinês	233,9
Holandês	74,1	Polonês	198,8
Total			27.281,6

Fonte: Casanova et al. (2024).

3.3.2 Bases Voltadas a Sotaques

Além dos corpora gerais de fala, alguns trabalhos vêm buscando construir bases especificamente voltadas à representação de sotaques e variações regionais.

O projeto Braccent, de Batista (2019), por exemplo, foi desenvolvido com foco na identificação automática de sotaques regionais brasileiros utilizando técnicas de aprendizado de máquina. A base é composta por gravações controladas de fala produzidas por falantes de diferentes regiões do Brasil, permitindo investigações relacionadas à caracterização acústica de sotaques.

O projeto Sotaque Brasileiro, descrito anteriormente, também se enquadra nessa categoria, dada sua coleta distribuída de frases padronizadas em diferentes regiões.

Esses corpora representam importantes iniciativas para mitigação do viés regional frequentemente observado em modelos treinados predominantemente com dados oriundos de regiões específicas do país, especialmente Sudeste e Sul do Brasil.

Além disso, estudos recentes em síntese de sotaque e transferência de estilo, como o de Zhou et al. (2023) sobre síntese com sotaque em cenários de dados limitados, vêm demonstrando a importância de bases multi-sotaque para tarefas de adaptação de locutor e síntese regionalizada. Em muitos casos, pequenas quantidades de dados específicos do sotaque alvo

podem ser utilizadas em conjunto com modelos pré-treinados de grande escala para induzir características regionais perceptíveis na fala sintetizada.

3.3.3 Limitações Identificadas na Literatura

Apesar dos avanços observados na disponibilização de corpora para o português brasileiro, a literatura ainda apresenta limitações relevantes relacionadas à representação de sotaques regionais em sistemas de síntese de fala.

Grande parte das bases públicas disponíveis possui volume reduzido de dados para determinadas regiões do país, especialmente quando consideradas aplicações envolvendo aprendizagem profunda e síntese de fala por redes neurais. As bases Sotaque Brasileiro e Braccent ilustram essa escassez: delas foi possível aproveitar, respectivamente, apenas 4 e 64 amostras associadas ao estado da Paraíba, somando pouco mais de 14 minutos de fala (Seção 5.2).

Além disso, diversos corpora foram originalmente desenvolvidos para tarefas de reconhecimento automático de fala ou análise sociolinguística, a exemplo do TAGARELA, derivado automaticamente de podcasts, e do CoLingPB, constituído por entrevistas sociolinguísticas e detalhado no Capítulo 4, não necessariamente apresentando características ideais para treinamento de sistemas TTS, como gravações limpas, segmentação adequada e consistência textual.

Outra limitação observada refere-se à escassez de trabalhos explicitamente voltados à modelagem do sotaque paraibano em sistemas de síntese de fala. Embora existam iniciativas relacionadas à diversidade regional do português brasileiro, tais como o Braccent e o Sotaque Brasileiro, não foram identificados estudos específicos envolvendo treinamento, adaptação ou avaliação sistemática de modelos TTS direcionados à reprodução desse sotaque regional.

Essas limitações evidenciam a necessidade de estratégias complementares de coleta, curadoria e adaptação de dados, bem como de investigações sobre adaptação de modelos abertos e avaliação perceptual da síntese de fala regionalizada em cenários de recursos limitados, no contexto do português brasileiro.

3.4 Síntese da Revisão Bibliográfica

A revisão bibliográfica permitiu identificar os principais avanços recentes em síntese de fala baseada em aprendizagem profunda e as estratégias utilizadas para modelagem de sotaques e adaptação regional em sistemas TTS. Arquiteturas modernas vêm possibilitando síntese com elevada naturalidade, além de clonagem de voz, síntese zero-shot e adaptação multilíngue com quantidades reduzidas de dados, e verificou-se a crescente disponibilidade de modelos abertos e corpora públicos para o português brasileiro. Persistem, contudo, as limitações de representação de sotaques regionais discutidas na Seção 3.3.3, em especial a ausência de trabalhos voltados à adaptação e à avaliação sistemática de modelos TTS para o sotaque paraibano, lacuna que motiva diretamente esta dissertação. A Tabela 3.2 sintetiza os principais trabalhos analisados. No próximo capítulo é descrita a metodologia adotada para endereçar essa lacuna, incluindo a construção do corpus, as estratégias de adaptação investigadas e os protocolos de avaliação.

3.4.1 Quadro Comparativo dos Trabalhos

A Tabela 3.2 apresenta um resumo comparativo de trabalhos, modelos e corpora discutidos ao longo desta revisão bibliográfica.

Tabela 3.2: Quadro comparativo de trabalhos relacionados à síntese de fala, corpora linguísticos e modelagem de sotaques.

Trabalho	**Ano**	**Contribuição Principal**
WaveNet (van den Oord et al., 2016)	2016	Introdução da síntese neural autorregressiva com geração direta da waveform de áudio.
Tacotron 2 (Shen et al., 2018)	2018	Conversão ponta a ponta de texto em espectrogramas mel utilizando mecanismos de atenção.

Trabalho	**Ano**	**Contribuição Principal**
FastSpeech (Ren et al., 2019)	2019	Síntese não autorregressiva com inferência paralela e maior estabilidade de alinhamento.
Grad-TTS (Popov et al., 2021)	2021	Aplicação de modelos de difusão para geração de espectrogramas mel com elevada qualidade perceptiva.
XTTS (Casanova et al., 2024)	2024	Síntese multilíngue zero-shot com clonagem de voz utilizando pequenas amostras de áudio.
Qwen3-TTS (Hu et al., 2026)	2026	Modelo TTS multimodal de grande escala com capacidades avançadas de síntese expressiva e clonagem de voz multilíngue.
F5-TTS (Chen et al., 2025)	2025	Modelo open-source voltado à adaptação de locutor e síntese multilíngue baseada em Flow Matching.
Orpheus TTS + Unsloth (Canopy Labs, 2025; Han et al., 2023)	2025	Estratégia de ajuste fino utilizando LoRA e otimizações de memória para adaptação eficiente de modelos TTS com menor custo computacional.
Braccent (Batista, 2019)	2019	Base de dados e sistema voltados à identificação automática de sotaques regionais brasileiros.
Sotaque Brasileiro (Milan, 2021)	2021	Corpus colaborativo para representação da diversidade regional da fala brasileira.
CoLingPB (Stein et al., 2013)	2013	Corpus linguístico composto por entrevistas de falantes paraibanos voltado a estudos sociolinguísticos e fonéticos.

Trabalho	**Ano**	**Contribuição Principal**
Zhou et al. (Zhou et al., 2023)	2023	Investigação de transferência de sotaque utilizando modelagem fonética e prosódica em sistemas TTS.

Fonte: elaborada pelo autor (2026).

Capítulo 4

Metodologia

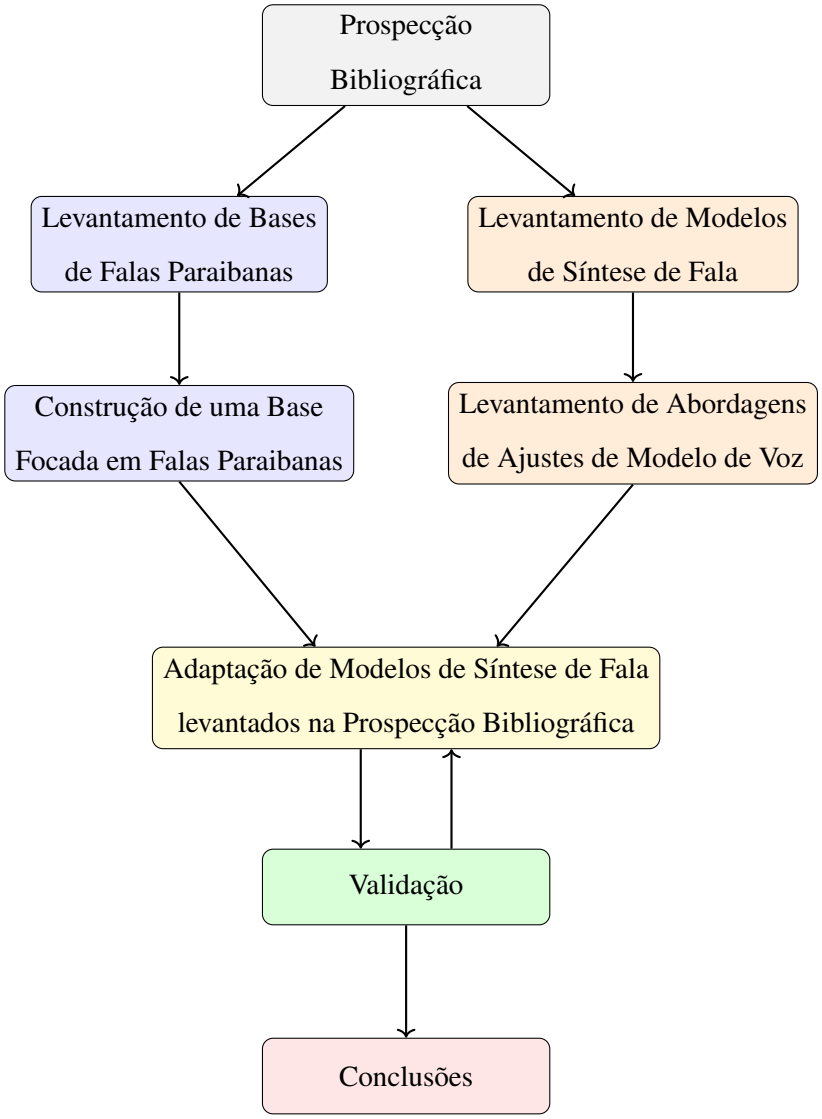
4.1 Visão Geral da Pesquisa

Este capítulo apresenta a metodologia adotada nesta pesquisa, de caráter experimental e aplicado, cujo objetivo é investigar a adaptação de modelos de síntese de fala para a reprodução do sotaque paraibano. O processo é organizado em etapas sequenciais e complementares, que vão desde o levantamento teórico até a validação experimental dos modelos adaptados.

Conforme apresentado no capítulo anterior, a primeira etapa metodológica foi uma prospecção bibliográfica para identificar modelos de síntese de fala, bases de dados disponíveis e técnicas de adaptação voltadas a sotaques regionais, seguida do levantamento e do teste desses modelos quanto à sua adequação aos objetivos da pesquisa. Em seguida, foram investigadas bases de fala em português brasileiro com variações regionais associadas ao sotaque paraibano, a partir das quais se construiu o corpus FALA_PB, integrando bases existentes, pré-processando corpora linguísticos e coletando novas amostras.

Os cinco modelos selecionados (XTTS, F5-TTS, Orpheus TTS, Fish Speech e Qwen3-TTS, detalhados na Seção 4.3.2) foram então adaptados sobre o FALA_PB segundo três estratégias de esforço computacional crescente: clonagem de fala a partir de áudios de referência, ajuste fino de *checkpoints* pré-treinados e treinamento a partir de pesos aleatórios. Os modelos resultantes foram avaliados por métricas automáticas de qualidade e inteligibilidade e por avaliação perceptiva com ouvintes humanos. Os resultados obtidos subsidiam as conclusões do trabalho. A visão geral dessas etapas é apresentada na Figura 4.1.

Figura 4.1: Etapas da metodologia adotada para adaptação de modelos de síntese de fala ao sotaque paraibano.



Fonte: elaborado pelo autor (2026).

4.2 Questões de Pesquisa

As questões de pesquisa e as hipóteses que orientam este trabalho foram apresentadas de forma sintética na Introdução (Seções 1.4 e 1.5). Esta seção as retoma em detalhe, associando a cada questão a hipótese correspondente e os procedimentos experimentais adotados para verificá-la.

- Questão de Pesquisa 1: Como construir uma base de dados de fala voltada ao sotaque

paraibano adequada para treinamento e adaptação de modelos de síntese de fala?

Hipótese: A integração de bases existentes, associada ao pré-processamento de corpora linguísticos e à coleta colaborativa de novas amostras de fala, possibilita a construção de uma base de dados com diversidade e qualidade suficientes para adaptação de modelos TTS ao sotaque paraibano.

- Questão de Pesquisa 2: Quais estratégias de adaptação são viáveis para a incorporação do sotaque paraibano em modelos de síntese de fala, considerando as restrições de recursos computacionais e de dados disponíveis?

  Hipótese: Estratégias com diferentes níveis de esforço computacional e quantidade de dados necessários apresentam comportamentos distintos quanto à capacidade de reproduzir características associadas ao sotaque paraibano. As estratégias investigadas estão enumeradas em ordem crescente de esforço necessário, considerando que a clonagem de fala representa o menor nível de adaptação por utilizar abordagens do tipo zero-shot. O ajuste fino de modelos previamente treinados demanda menor quantidade de dados e menos épocas de treinamento quando comparado ao treinamento completo de modelos, enquanto o treinamento a partir de pesos aleatórios representa o maior custo computacional e necessidade de dados. Nesta pesquisa, a viabilidade das três estratégias é investigada experimentalmente: a clonagem de voz e o treinamento a partir de pesos aleatórios são examinados em caráter exploratório, enquanto a estratégia identificada como mais promissora é validada sistematicamente pelas avaliações automática e perceptiva.

  1. clonagem de fala;
  2. ajuste fino de modelos previamente treinados;
  3. treinamento completo de modelos utilizando exclusivamente a base FALA_PB.

- Questão de Pesquisa 3: Os ajustes realizados nos modelos de síntese de fala impactam negativamente a qualidade e inteligibilidade das falas sintetizadas?

Hipótese: Estratégias de adaptação de sotaque podem introduzir alterações na qualidade e inteligibilidade das falas sintetizadas. Para investigação dessa hipótese serão utilizadas métricas automáticas, incluindo preditores de MOS e métricas de similaridade textual baseadas em funções de distância aplicadas sobre transcrições automáticas das falas geradas.

- Questão de Pesquisa 4: Os ajustes realizados nos modelos de síntese de fala conseguem produzir percepção consistente do sotaque paraibano nas falas sintetizadas?

  Hipótese: Os modelos adaptados utilizando a base FALA_PB conseguem reproduzir características associadas ao sotaque paraibano, sendo essa percepção validada por meio de avaliações realizadas com ouvintes humanos.

4.3 Materiais

Esta seção apresenta os principais materiais utilizados ao longo da pesquisa, incluindo as bases de dados empregadas na construção do corpus FALA_PB, os modelos de síntese de fala investigados, bem como a infraestrutura computacional e ferramentas utilizadas durante os experimentos.

4.3.1 Bases de Dados Utilizadas

Para construção do corpus FALA_PB foram utilizadas bases de dados contendo falas em português brasileiro, com foco em materiais relacionados ao sotaque paraibano e demais variações regionais. As bases utilizadas possuem diferentes características em relação à quantidade de falantes, duração dos áudios, formato de disponibilização e necessidade de pré-processamento. A caracterização quantitativa e a composição final do corpus integrado são apresentadas na Seção 5.2. O uso das bases Sotaque Brasileiro, Braccent e do CoLingPB contou com autorização dos respectivos responsáveis.

Além das bases já existentes, também foram coletadas novas amostras de fala por meio de uma coleta colaborativa realizada com falantes paraibanos, incluindo leitura de trechos de *Vidas Secas* (Ramos, 1938) e gravações enviadas via WhatsApp, permitindo ampliar a diversidade e quantidade de dados disponíveis para os experimentos realizados nesta pesquisa.

Sotaque Brasileiro

A base de dados Sotaque Brasileiro (Milan, 2021) consiste em um conjunto de falas organizadas por estados brasileiros, feito por meio de coleta colaborativa, permitindo a identificação de diferentes regionalismos presentes no português brasileiro. A base apresenta estrutura adequada para utilização em tarefas de síntese de fala, contendo áudios segmentados e organizados de forma relativamente padronizada. Para esta pesquisa, foram selecionadas amostras associadas ao estado da Paraíba, utilizadas como complemento na construção do corpus *FALA_PB*.

Braccent

A base de dados Braccent (Batista, 2019) foi desenvolvida com foco em estudos relacionados a sotaques do português brasileiro e reúne falas organizadas por diferentes regiões e estados do país. A base apresenta contribuição relevante para pesquisas envolvendo classificação e análise de sotaques regionais, além de fornecer materiais potencialmente aplicáveis em tarefas de síntese de fala. Para esta pesquisa, foram selecionadas amostras relacionadas ao sotaque paraibano, integradas como complemento ao corpus *FALA_PB*.

Base Colaborativa

A coleta colaborativa reuniu dois subconjuntos: frases enviadas por participantes via WhatsApp e leitura de *Vidas Secas* em ambiente controlado. O protocolo está na Seção 4.4.2 e os totais consolidados, na Seção 5.2.3.

Corpus Linguístico da Paraíba (CoLingPB)

O Corpus Linguístico da Paraíba (CoLingPB) (Stein et al., 2013) consiste em uma base de dados voltada ao estudo linguístico de falantes paraibanos, contendo um total de 324 entrevistas coletadas em diferentes regiões do estado da Paraíba. O corpus foi desenvolvido no âmbito da Universidade Federal da Paraíba (UFPB), sob coordenação do professor Cirineu Cecote Stein, e apresenta relevância significativa para pesquisas relacionadas à variação linguística e regionalismos do português brasileiro.

As entrevistas originais são identificadas por um padrão de nomenclatura nos arquivos

.wav que codifica município, faixa etária, zona urbana/rural, sexo e número do informante (detalhado na Seção 4.4).

Diferentemente das bases Sotaque Brasileiro e Braccent, o CoLingPB não foi originalmente estruturado para utilização em tarefas de síntese ou reconhecimento de fala. Os áudios disponibilizados consistem majoritariamente em entrevistas contendo tanto a fala do entrevistador quanto a fala do participante entrevistado, além de variações de qualidade acústica entre os registros disponíveis.

Apesar dessas limitações, o corpus apresenta grande relevância para esta pesquisa devido à quantidade de falas associadas diretamente ao sotaque paraibano, tornando-se uma das principais fontes de dados do FALA_PB. Os procedimentos de extração, segmentação e filtragem necessários para adequá-lo ao formato esperado pelos modelos são detalhados na Seção 4.4.

4.3.2 Modelos de Síntese de Fala Investigados

Os modelos investigados nesta pesquisa foram selecionados considerando critérios relacionados à disponibilidade de código aberto, possibilidade de execução local dos experimentos, suporte a clonagem de fala e existência de checkpoints compatíveis com português brasileiro, mesmo que disponibilizados de forma não oficial por comunidades e usuários em plataformas como Hugging Face.

Além disso, buscou-se selecionar modelos recentes e amplamente utilizados pela comunidade de síntese de fala, permitindo avaliar diferentes estratégias de adaptação de sotaque utilizando arquiteturas modernas de TTS.

Dessa forma, foram selecionados os seguintes modelos para os experimentos realizados nesta dissertação:

1. F5-TTS;
2. XTTS;
3. Orpheus TTS;
4. Fish Speech;
5. Qwen3-TTS;

4.3.3 Infraestrutura Computacional

Os experimentos realizados nesta pesquisa utilizaram uma infraestrutura computacional composta por processamento em GPU para treinamento e adaptação dos modelos de síntese de fala.

A Tabela 4.1 apresenta a configuração utilizada durante os experimentos.

Tabela 4.1: Infraestrutura computacional utilizada nos experimentos.

Componente	**Especificação**
Sistema Operacional	Windows 11 (AMD64)
CPU	AMD, 6 núcleos / 12 threads lógicos
Memória RAM	32 Gb
GPU	NVIDIA GeForce RTX 4060 Ti
Memória VRAM	16 GB
PyTorch	2.5.1+cu121
CUDA	12.1
Driver NVIDIA	596.21
Docker	Docker Engine 28.1.1

Fonte: elaborada pelo autor (2026).

4.3.4 Ferramentas e Bibliotecas Utilizadas

Durante o desenvolvimento desta pesquisa foram utilizadas diferentes ferramentas e bibliotecas voltadas ao processamento de áudio, treinamento de modelos de síntese de fala, manipulação de dados e execução dos experimentos computacionais.

Os experimentos envolvendo treinamento e adaptação dos modelos foram realizados principalmente utilizando a biblioteca PyTorch, juntamente com bibliotecas auxiliares presentes nos repositórios oficiais dos modelos investigados. Entre as principais bibliotecas utilizadas durante os experimentos destacam-se TorchAudio, Accelerate, Transformers e Gradio, amplamente empregadas nos repositórios dos modelos investigados nesta dissertação.

Para gerenciamento dos ambientes de execução e padronização dos experimentos foram utilizados contêineres Docker, permitindo maior reprodutibilidade dos ambientes de treinamento e inferência.

Durante as etapas de processamento de áudio foi utilizada a ferramenta FFmpeg para conversão, segmentação, padronização e manipulação dos arquivos de fala utilizados no corpus FALA_PB. Além disso, bibliotecas auxiliares de manipulação de áudio e processamento numérico também foram utilizadas durante as etapas de pré-processamento e filtragem das falas coletadas.

Para transcrição automática das falas foi utilizado o modelo Whisper (Radford et al., 2023), empregado principalmente nas etapas de avaliação automática e comparação textual entre as falas sintetizadas e os textos originalmente esperados.

Além disso, bibliotecas de visualização e análise de dados foram utilizadas para geração de tabelas, histogramas e gráficos relacionados à distribuição dos dados e aos resultados experimentais. Entre as principais ferramentas utilizadas destacam-se Matplotlib e bibliotecas auxiliares de visualização estatística empregadas ao longo da pesquisa.

4.4 Construção do Corpus FALA_PB

Esta seção apresenta os procedimentos realizados para construção do corpus FALA_PB, desenvolvido com o objetivo de reunir falas associadas ao sotaque paraibano para utilização nos experimentos de síntese de fala realizados nesta pesquisa.

A construção do corpus envolveu a integração de diferentes fontes de dados, incluindo bases previamente existentes e coletas colaborativas realizadas ao longo da pesquisa. Além da obtenção dos áudios, também foram necessárias etapas adicionais de extração, segmentação, filtragem, padronização e controle de qualidade dos arquivos, principalmente devido às diferenças estruturais entre as bases utilizadas.

Em especial, o Corpus Linguístico da Paraíba (CoLingPB) demandou etapas adicionais de extração e segmentação, detalhadas a seguir.

As subseções seguintes detalham os procedimentos realizados durante as etapas de construção e tratamento do corpus FALA_PB.

4.4.1 Extração e Segmentação do CoLingPB

Como o CoLingPB é disponibilizado em entrevistas completas, foi necessário segmentá-lo em amostras individuais de fala. A partir das marcações temporais dos arquivos *.trs*, os

trechos de fala dos entrevistados foram recortados automaticamente e filtrados por qualidade acústica. O detalhamento do procedimento está no Apêndice B, e os volumes resultantes, na Seção 5.2.

4.4.2 Coleta Colaborativa de Falas

Além das bases previamente existentes, foi realizada uma coleta colaborativa de falas para ampliar o volume de dados associados ao sotaque paraibano no corpus *FALA_PB*. A iniciativa surgiu após a identificação do desbalanceamento regional nas bases públicas analisadas inicialmente, com pouca representação de falas nordestinas, e antecedeu a incorporação do CoLingPB ao projeto.

Foram adotadas duas estratégias complementares, detalhadas a seguir. A composição quantitativa final dos arquivos obtidos em cada uma está na Tabela 5.3 (resultados na Seção 5.2.3). A coleta colaborativa foi precedida de um Termo de Consentimento Livre e Esclarecido (TCLE), assim como a etapa de avaliação perceptiva.

Coleta Colaborativa via WhatsApp

Cada participante foi solicitado a gravar três frases independentes, com duração aproximada de dez segundos cada uma em leitura em velocidade normal. Os convites foram feitos a voluntários entre colegas universitários, amigos e familiares do pesquisador, nativos ou habituais falantes da região. As gravações foram enviadas pelo WhatsApp, preferencialmente a partir do próprio celular de cada participante e em ambiente silencioso quando possível.

As frases que deveriam ser lidas pelos participantes foram geradas de forma assistida por um modelo de linguagem (ChatGPT, modelo GPT-4o), a partir de um *prompt* estruturado que exige mensagem curta para WhatsApp, tom informal, três frases numeradas, uso exclusivamente acadêmico, ausência de identificação e de uso comercial, variedade temática e lexical e linguagem cotidiana. O prompt completo está no Apêndice A. Para cada novo participante, as três frases são geradas em execução separada, de modo a reduzir repetição temática entre convites distintos.

O fluxo de contato com os participantes voluntários seguiu etapas fixas: (i) apresentação do pesquisador e do objetivo da dissertação; (ii) mensagem de convite pedindo autorização

para encaminhar as frases; (iii) em caso de aceite, envio da mensagem com as três frases a serem lidas; (iv) em caso de recusa, agradecimento e encerramento do pedido.

Exemplos da mensagem de convite e de uma solicitação efetivamente enviada a um participante constam do Apêndice A.

Coleta Colaborativa com Leitura de *Vidas Secas*

Utilizou-se a obra *Vidas Secas* (Ramos, 1938), de Graciliano Ramos, em domínio público. Capítulos selecionados foram divididos em sentenças e lidos em voz alta, com gravação em sala silenciosa utilizando o microfone Arcano Tsion. Durante a leitura, as pausas curtas foram aproveitadas para segmentar automaticamente cada trecho falado em um arquivo separado; trechos sem fala audível suficiente foram descartados e os demais receberam identificador sequencial, com os intervalos de início e fim registrados. Esse procedimento produziu a maior parte do volume de amostras da coleta colaborativa (Tabela 5.3).

Organização e Pós-Processamento

Os áudios de ambos os subconjuntos foram armazenados com o texto lido correspondente e passaram pelas etapas de filtragem e controle de qualidade aplicadas às demais fontes do *FALA_PB*, além da padronização para o formato exigido pelos modelos (mono, 24 kHz, PCM 16 bits na versão consolidada do corpus). Foram desenvolvidas ferramentas auxiliares para apoiar gravação, organização e armazenamento.

Optou-se por não reportar o número de falantes distintos na coleta colaborativa, em especial no subconjunto via WhatsApp, para maximizar a privacidade dos voluntários da região e limitar riscos ligados à identificação por voz. As mensagens de convite deixavam claro o uso exclusivamente acadêmico, a ausência de identificação nominal e a não utilização comercial dos áudios; os arquivos foram tratados sem associação a nomes nos metadados divulgados.

4.4.3 Filtragem das Bases Sotaque Brasileiro e Braccent

As bases Sotaque Brasileiro e Braccent demandaram apenas um processo simples de filtragem dos áudios relacionados ao sotaque paraibano, uma vez que os dados já se encontravam previamente segmentados e organizados em arquivos individuais de fala. O volume de

amostras paraibanas identificado em cada base e a justificativa para uso complementar no *FALA_PB* são apresentados na Seção 5.2.

4.4.4 Filtragem e Controle de Qualidade

Após a integração e a segmentação das bases, foi necessário um procedimento de filtragem e controle de qualidade dos áudios, dada a heterogeneidade acústica das fontes: enquanto algumas amostras provinham de ambientes controlados, outras apresentavam ruído de fundo, reverberação, sons ambientes e diferenças de volume e captação. A análise manual identificou problemas como ruídos de veículos, pássaros e conversas paralelas, reverberação excessiva, distorções de microfone, pausas longas, cortes incompletos de fala e trechos de baixa inteligibilidade.

Para evitar a validação exclusivamente manual dos milhares de segmentos gerados, empregou-se o preditor *DNSMOS* (Reddy et al., 2022), que estima a qualidade diretamente do sinal de áudio, sem necessidade de referência, em quatro dimensões: *p808* (qualidade percebida geral), *sig* (qualidade da voz), *bak* (qualidade do ruído de fundo) e *ovrl* (qualidade geral). Adotou-se o escore *ovrl* como critério de corte, com limiar definido manualmente a partir da escuta de amostras (Seção 5.2). Essa triagem automática, combinada à escuta manual, fundamentou a remoção das amostras inadequadas (ruído intenso, baixa inteligibilidade, distorções severas, cortes incompletos ou ausência significativa de fala). A composição quantitativa do corpus após essa etapa é apresentada na Seção 5.2.

Padronização dos Arquivos

Após as etapas de coleta, extração e filtragem dos áudios, foi necessário realizar a padronização dos arquivos presentes no corpus FALA_PB para adequação ao formato esperado pelos modelos de síntese de fala investigados nesta pesquisa.

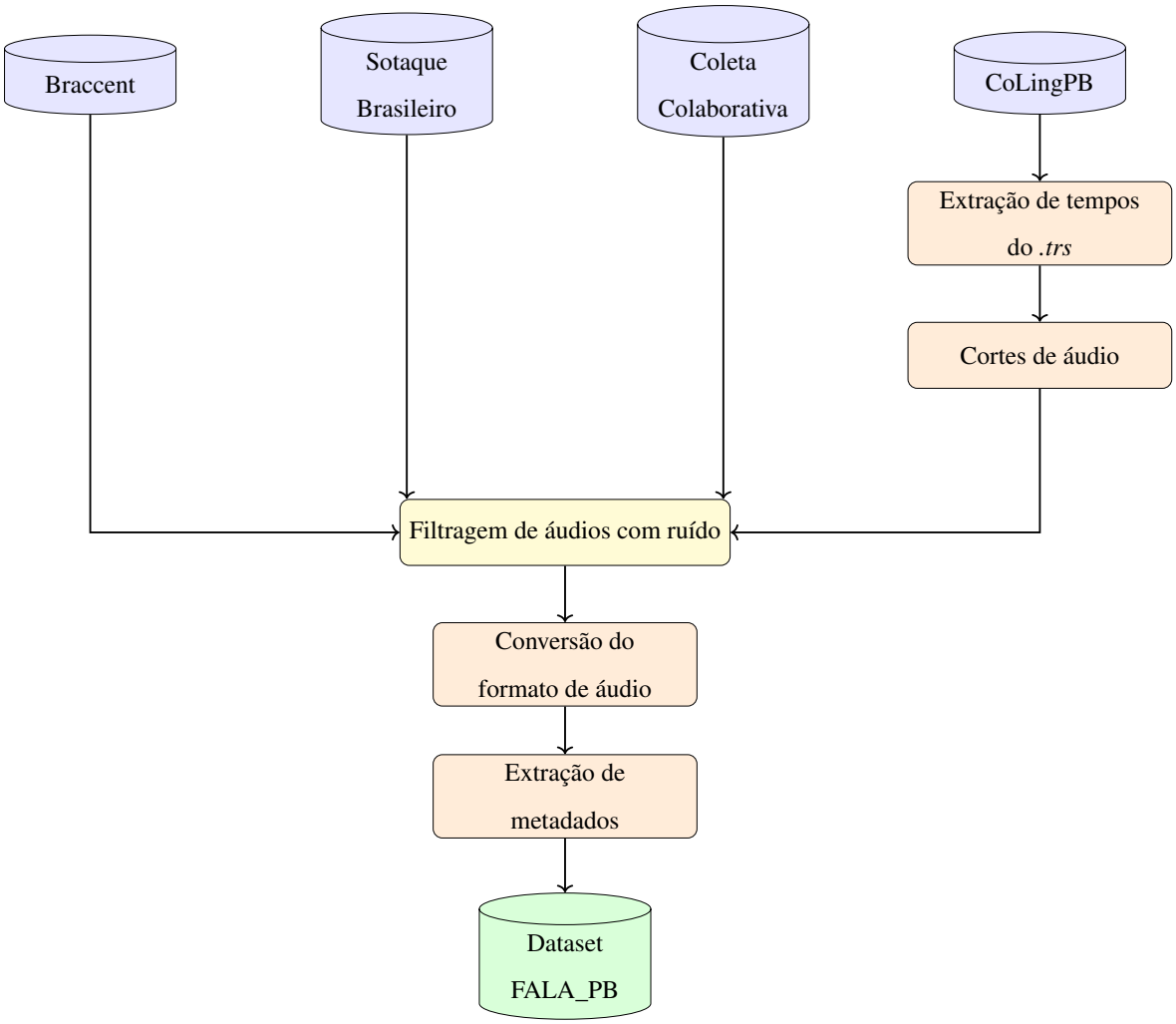
Considerando os requisitos dos modelos, os arquivos foram convertidos para o formato *.wav*, mono, com taxa de amostragem de 24 kHz e codificação PCM de 16 bits. Essa etapa ocorre após segmentação e filtragem; em particular, os segmentos do CoLingPB gerados a 16 kHz são reamostrados nesta fase. Os totais consolidados do *FALA_PB* reportados no Capítulo 5 referem-se a essa versão padronizada.

O processo de conversão e padronização foi realizado utilizando a ferramenta FFmpeg, permitindo uniformizar os diferentes formatos originalmente presentes nas bases utilizadas e nas gravações colaborativas. Essa etapa também possibilitou reduzir incompatibilidades entre os arquivos durante os experimentos de treinamento, inferência e avaliação dos modelos investigados.

4.4.5 Integração de Bases Existentes

A integração das fontes envolveu tanto a incorporação direta de amostras já segmentadas (Sotaque Brasileiro e Braccent) quanto o processamento adicional de entrevistas (CoLingPB) e de gravações da coleta colaborativa. A Figura 4.2 ilustra o fluxo de processamento das bases preexistentes, da extração e organização dos áudios até a geração dos segmentos padronizados.

Figura 4.2: Fluxo de processamento das bases de fala para construção do corpus FALA_PB.



Fonte: elaborado pelo autor (2026).

Após integração, filtragem e padronização, os sinais de áudio e metadados textuais foram consolidados no corpus final *FALA_PB*. A composição quantitativa do corpus, as limitações identificadas e a discussão da Questão de Pesquisa 1 encontram-se na Seção 5.2.

4.5 Estratégias de Adaptação dos Modelos

Esta seção apresenta as estratégias e os protocolos experimentais adotados para adaptação dos modelos de síntese de fala investigados nesta dissertação, utilizando o corpus *FALA_PB*. Os resultados de cada abordagem são reportados no Capítulo 5.

Foram investigadas três abordagens, em ordem crescente de esforço computacional e

de dependência de dados: (i) clonagem de fala; (ii) ajuste fino de modelos pré-treinados; (iii) treinamento completo a partir de pesos aleatórios. As subseções seguintes detalham os procedimentos, modelos e critérios de cada etapa; a comparação entre estratégias é feita nos resultados (Seção 5.3).

4.5.1 Clonagem de Fala

Como primeira estratégia de adaptação investigada nesta dissertação, foram definidos experimentos exploratórios de clonagem de fala (*voice cloning*), abordagem de menor esforço computacional entre as previstas na Questão de Pesquisa 2, por não exigir atualização de pesos do modelo em cenários *zero-shot*.

O objetivo metodológico dessa etapa foi verificar se modelos com suporte a clonagem conseguiriam reproduzir, a partir de poucos segundos de áudio de referência, características associadas ao sotaque paraibano, e não apenas o timbre do locutor, em textos de síntese distintos dos trechos utilizados como referência.

Modelos e Material de Referência

Foram considerados, entre os modelos selecionados para a pesquisa (Seção 4.3.2), aqueles que disponibilizam inferência com condicionamento por áudio de referência sem treinamento adicional, em especial arquiteturas multilíngues com suporte ao português brasileiro. Os experimentos foram conduzidos com o XTTS e o F5-TTS, ambos com suporte nativo a clonagem *zero-shot* em português brasileiro. As amostras de referência foram extraídas do corpus *FALA_PB*, priorizando falas de locutores paraibanos com conteúdo representativo do sotaque alvo.

Protocolo Experimental

Para cada modelo avaliado nesta etapa, foram preparados trechos de referência com durações aproximadas de 10, 30 e 60 segundos, permitindo comparar o efeito do volume de áudio de condicionamento sobre a síntese subsequente. A partir de cada referência, foram geradas amostras sintéticas para textos de entrada distintos dos conteúdos presentes no áudio de referência, de modo a simular uso real de TTS (texto arbitrário) e não apenas repetição ou

paráfrase mínima do trecho clonado.

A avaliação nessa fase foi predominantemente qualitativa, por escuta manual, com atenção a:

- preservação do timbre e identidade vocal do locutor de referência;
- presença perceptível de traços de sotaque paraibano na fala sintetizada;
- degradação da inteligibilidade ou naturalidade quando o texto de síntese se afasta do conteúdo da referência.

Os resultados observados, as limitações identificadas e a decisão metodológica sobre o papel da clonagem nas etapas seguintes são apresentados na Seção 5.3.1.

4.5.2 Ajuste Fino de Modelos Pré-Treinados

Na sequência dos experimentos de clonagem (Seção 5.3.1), foi definida a etapa de ajuste fino (*fine-tuning*) sobre *checkpoints* já treinados, utilizando o corpus *FALA_PB*. O objetivo foi adaptar os pesos dos modelos para incorporar características do sotaque paraibano, com menor custo computacional e menor volume de dados do que o treinamento completo a partir de pesos aleatórios.

Foram treinadas diferentes arquiteturas de síntese de fala (Seção 4.3.2), comparando procedimentos de adaptação, estabilidade do treino e viabilidade na infraestrutura disponível. Os resultados por modelo estão na Seção 5.3.2.

Critérios de Parada e Monitoramento

Os treinos foram executados na infraestrutura computacional pessoal descrita anteriormente, o que limitou tempo de execução e número de épocas.

Inicialmente, considerou-se encerrar o treinamento com base na curva de perda. Em modelos já pré-treinados, porém, essa métrica nem sempre indica convergência de forma clara. Adotou-se, então, parada por número de épocas e tempo total: no máximo três épocas completas; se o treino ultrapassasse cerca de 24 horas, encerramento após a primeira época.

Durante o treino foram registrados tempo de processamento e perda por passo e por época (quando aplicável). Ao final de cada época, geraram-se amostras sintéticas para escuta manual, observando inteligibilidade, estabilidade e indícios de sotaque paraibano.

Amostras de Acompanhamento Qualitativo

Ao fim de cada época avaliada, sintetizaram-se cinco frases fixas para monitoramento preliminar:

- "Hoje acordei cedo, preparei um café forte e organizei a mesa para estudar com calma."
- "Enquanto o trem passava devagar, uma criança sorria e apontava para as nuvens alaranjadas."
- "O pesquisador analisou os dados, escreveu um relatório objetivo e compartilhou as conclusões com a equipe."
- "Estamos construindo uma rotina mais saudável, caminhando no bairro e cozinhando alimentos frescos todos os dias."
- "A bibliotecária catalogou romances, dicionários e biografias, mantendo cada prateleira limpa e bem sinalizada."

Checkpoints

Os pesos intermediários e finais foram armazenados como *checkpoints*, para retomada do treino, comparação entre épocas e síntese nas etapas de avaliação automática e perceptual.

XTTS

O XTTS foi incluído por ser multilíngue, com suporte ao português brasileiro e código aberto para treino e inferência locais. Como ponto de partida, utilizou-se o checkpoint multilíngue padrão distribuído com o próprio modelo (XTTS-v2)[1], que já inclui suporte nativo ao português brasileiro, sem necessidade de checkpoint intermediário de terceiros. O ajuste fino usou o corpus *FALA_PB* por três épocas, conforme os critérios de parada acima. O treino

[1] `https://huggingface.co/coqui/XTTS-v2`

seguiu a configuração padrão de ajuste fino disponibilizada no repositório do modelo, sem alteração dos hiperparâmetros originais; por isso, não se reporta uma tabela específica de hiperparâmetros para o XTTS.

F5-TTS

O F5-TTS foi incluído por ser arquitetura recente de síntese neural, de código aberto e com receitas de treino no repositório do projeto. O *checkpoint* inicial foi `firstpixel/F5-TTS-pt-br`[2], adaptado com *FALA_PB* por três épocas. A Tabela 4.2 resume os hiperparâmetros.

Tabela 4.2: Hiperparâmetros do ajuste fino do F5-TTS.

Hiperparâmetro	**Valor**
Checkpoint utilizado	firstpixel/F5-TTS-pt-br
Batches por época	17.642
Quantidade de épocas	3
Tipo de batch	Frame
Batch size por GPU	1000
Máximo de amostras simultâneas	64
Gradiente acumulado	1
Taxa de aprendizado	1×10^{-5}

Fonte: elaborada pelo autor (2026).

Orpheus TTS

O Orpheus TTS foi ajustado com LoRA via biblioteca Unsloth, a partir do checkpoint `ArtooDtoo/Orpheus_PTBR_FT_Unsloth_Chk-1800`[3] e do corpus *FALA_PB*.

O pipeline exige etapa intermediária de processamento dos áudios antes do treino, o que aumenta o tempo total. Por isso, limitou-se o treino a apenas uma época. A Tabela 4.3 lista os hiperparâmetros.

[2] `https://huggingface.co/firstpixel/F5-TTS-pt-br`

[3] `https://huggingface.co/ArtooDtoo/Orpheus_PTBR_FT_Unsloth_Chk-1800`

Tabela 4.3: Hiperparâmetros do ajuste fino do Orpheus TTS.

Hiperparâmetro	**Valor**
Taxa de aprendizado	1×10^{-5}
Batch size por dispositivo	1
Gradiente acumulado	4
Warmup steps	5
Tamanho máximo de sequência	1024
Taxa de amostragem dos áudios	24 kHz
Seed de aleatoriedade	3407

Fonte: elaborada pelo autor (2026).

Fish Speech

O Fish Speech foi ajustado com LoRA, a partir do *checkpoint* multilíngue padrão distribuído no repositório oficial do modelo, que possui suporte nativo ao português brasileiro. O treino exige pré-processamento e extração de representações intermediárias dos áudios. Devido ao custo computacional, executou-se uma época de ajuste fino. A Tabela 4.4 resume a configuração.

Tabela 4.4: Hiperparâmetros do ajuste fino do Fish Speech.

Hiperparâmetro	**Valor**
Tipo de adaptação	LoRA
Rank LoRA (r)	8
LoRA Alpha	16
Taxa de aprendizado	1×10^{-4}
Batch size	4
Precisão numérica	bf16
Warmup steps	10
Gradient clipping	1,0
Tamanho máximo de sequência	4096

Fonte: elaborada pelo autor (2026).

Qwen3-TTS

O Qwen3-TTS (Hu et al., 2026) foi incluído por síntese multilíngue expressiva, suporte a clonagem de voz e relevância para adaptação de estilo e locutor, além de ser uma das referências mais recentes que se buscou adotar. A Tabela 4.5 resume os hiperparâmetros de treino. A taxa de aprendizado corresponde ao valor de referência disponibilizado pelos autores do modelo (Hu et al., 2026), mantido sem alteração nesta pesquisa; a discussão do seu efeito sobre o resultado do ajuste encontra-se na Seção 5.3.5.

Entre os ajustes finos desta dissertação, o Qwen3-TTS foi um dos mais caros computacionalmente em tempo de GPU: a época única demandou cerca de quatro dias, três horas e vinte e oito minutos, acima do limite operacional de um dia adotado para outros modelos, o que reforçou a execução de apenas uma época.

Tabela 4.5: Hiperparâmetros do ajuste fino do Qwen3-TTS.

Hiperparâmetro	**Valor**
Checkpoint utilizado	Qwen3-TTS-12Hz-1.7B-Base
Quantidade de épocas	1
Batch size por GPU	2
Gradiente acumulado	4
Batch efetivo	8
Taxa de aprendizado	2×10^{-6}
Precisão numérica	bf16

Fonte: elaborada pelo autor (2026).

4.5.3 Treinamento a partir de Pesos Aleatórios

Como terceira estratégia da Questão de Pesquisa 2, foi definido o treinamento completo do F5-TTS a partir de pesos aleatórios, sem *checkpoint* pré-treinado. A motivação veio do ajuste fino (Seção 5.3.2): os pesos mudaram, mas a síntese permaneceu praticamente igual à do modelo base. O treinamento a partir de pesos aleatórios testa se o sotaque paraibano se incorpora melhor quando a rede aprende quase só com o *FALA_PB*.

Nesta dissertação, o treinamento a partir de pesos aleatórios funciona como estudo de caso da estratégia mais custosa da Questão de Pesquisa 2, e não como retreinamento de

todas as arquiteturas selecionadas. Mesmo com desfecho negativo na síntese (Seção 5.3.3), o procedimento é documentado para transparência e para contraste com clonagem e ajuste fino.

Motivação para Utilizar o F5-TTS no Treinamento com Pesos Aleatórios

Dois motivos levaram à escolha do F5-TTS. O primeiro é que foi o caso sem resultado positivo no ajuste fino: os pesos foram alterados, mas a síntese ficou praticamente igual à do *checkpoint* pré-treinado. Retreinar a partir de pesos aleatórios verifica se o impasse estava no *checkpoint* e na dinâmica do *fine-tuning*, e não na impossibilidade de o modelo aprender padrões do *FALA_PB*.

O segundo é prático: entre os repositórios observados nesta dissertação, o do F5-TTS ofereceu o fluxo mais simples para treino com pesos aleatórios, com receitas documentadas, compatibilidade com o formato de dados já usado no ajuste fino e interface Gradio para configurar e acompanhar o treino. Orpheus TTS e Fish Speech exigiriam etapas extras de pré-processamento, além de tempo e disco indisponíveis na infraestrutura utilizada.

Modelo, Dados e Critério de Parada

O treinamento partiu de pesos aleatórios, sem checkpoint pré-treinado nem ajuste fino prévio, sobre o mesmo corpus das rodadas de ajuste fino.

O limite operacional de GPU foi de cerca de dez dias de execução contínua, intervalo estimado para comportar cerca de 150 épocas. Os hiperparâmetros adotados estão na Tabela 4.6, e registraram-se perda de treino e *eval loss* por época. Para teste informal, previu-se a síntese das cinco frases de acompanhamento da Seção 4.5.2.

Tabela 4.6: Hiperparâmetros do treinamento do F5-TTS a partir de pesos aleatórios.

Hiperparâmetro	**Valor**
Inicialização	Pesos aleatórios (sem checkpoint pré-treinado)
Quantidade de épocas	150
Taxa de aprendizado	$7{,}5 \times 10^{-5}$
Batch	1000 frames/GPU, máx. 64 amostras

Fonte: elaborada pelo autor (2026).

Os resultados quantitativos, a curva de perda, a discussão do desfecho e a exclusão deste checkpoint das avaliações perceptivas planejadas estão na Seção 5.3.3.

4.6 Avaliação Automática dos Modelos

Esta seção descreve o protocolo de avaliação automática conduzido para a Questão de Pesquisa 3: verificar se os ajustes nos modelos de síntese de fala afetam a qualidade e a inteligibilidade das falas sintetizadas. A fundamentação sobre preditores de MOS e similaridade textual está no Capítulo 2; aqui define-se o conjunto de métricas e o fluxo de cálculo aplicados aos checkpoints gerados nesta dissertação.

Os resultados agregados por modelo e estratégia são reportados na Seção 5.4 do Capítulo 5. A avaliação perceptiva com ouvintes (comparações perceptivas de sotaque e de clareza/naturalidade) permanece na Seção 4.7.

4.6.1 Visão Geral e Amostras Avaliadas

Foram sintetizadas falas a partir de um conjunto fixo de 100 sentenças de teste em português brasileiro, distintas das usadas no treinamento e das cinco frases de acompanhamento do ajuste fino (Seção 4.5.2), para cada modelo e estratégia de adaptação comparável (clonagem, ajuste fino, treinamento a partir de pesos aleatórios). Cada sentença foi sintetizada em duas condições (*checkpoint* original e *checkpoint* ajustado); para os três modelos que chegaram a esta etapa, isso totaliza $3 \times 2 \times 100 = 600$ áudios. Dessa forma, cada sentença contribui com um único par original–ajustado, e os testes pareados incidem sobre 100 pares por modelo. Sobre cada áudio sintetizado foram calculadas, de forma reprodutível:

- dimensões de qualidade calculadas por um preditor de MOS;
- distância de Levenshtein normalizada entre o texto enviado ao TTS e a transcrição obtida pelo Whisper, como indicador lexical de preservação do conteúdo.

A Tabela 4.7 lista as métricas em português.

Tabela 4.7: Métricas automáticas calculadas sobre cada amostra de fala sintetizada.

Métrica	**Descrição**
Preditor de MOS	
Qualidade geral calculada	Estimativa da qualidade perceptual global do áudio.
Ruído percebido calculado	Nível de ruído percebido no sinal.
Descontinuidade calculada	Presença de cortes ou quebras audíveis.
Coloração calculada	Alterações tonais e timbrais percebidas.
Intensidade sonora calculada	Equilíbrio de volume e intensidade sonora.
Similaridade textual (STT + Levenshtein)	
Distância de Levenshtein (normalizada)	Diferença proporcional em relação ao tamanho da referência; valores próximos de zero indicam maior correspondência.

Os preditores de MOS adotados foram o *NISQA* e o *DNSMOS*; as dimensões reportadas seguem os nomes da tabela (detalhamento na Seção 5.4).

Fonte: elaborada pelo autor (2026), com base em Mittag et al. (2021) e Reddy et al. (2022).

A Tabela 4.8 sintetiza as várias metodologias de avaliação de modelos para síntese de fala adotadas nesta dissertação.

Tabela 4.8: Comparação entre diferentes metodologias de avaliação em síntese de fala.

Métrica	**Custo**	**Observação**
MOS (*Mean Opinion Score*)	Alto	Requer participação de avaliadores humanos. Considerada uma das principais métricas subjetivas para avaliação de naturalidade e qualidade perceptiva da fala sintetizada.
Preditor de MOS (qualidade calculada)	Baixo	Preditor não intrusivo de MOS aplicado ao áudio sintetizado (implementação configurável), sem avaliadores humanos. Adequado para análises preliminares e comparações rápidas.
Levenshtein + STT	Baixo	Avalia indiretamente a inteligibilidade por meio da comparação entre o texto original e a transcrição obtida via reconhecimento automático de fala.

Fonte: elaborada pelo autor (2026).

4.6.2 Preditores de MOS

Utiliza-se um preditor de MOS que estima aspectos de qualidade perceptiva diretamente do áudio sintetizado. Os resultados são registrados com o sufixo *calculada* ou *calculado* no nome da métrica, para distinguir essas estimativas do MOS obtido com ouvintes humanos (Seção 4.7).

Adotaram-se as implementações *NISQA* (Mittag et al., 2021) e *DNSMOS* (Reddy et al., 2022). As dimensões reportadas incluem qualidade geral, ruído percebido, descontinuidade, coloração e intensidade sonora (Tabela 4.7).

Os valores das métricas acima foram comparados de forma relativa entre modelos, *checkpoints* e estratégias (mesmos textos de teste, mesmo pipeline). Em geral, escores mais altos na escala interna do preditor indicam melhor qualidade perceptiva estimada.

4.6.3 Transcrição Automática das Falas

Para a análise lexical, cada áudio sintetizado foi transcrito com o modelo Whisper (Radford et al., 2023), em sua variante *small*, com idioma fixo em português e parâmetros de decodificação constantes entre execuções.

4.6.4 Similaridade Textual: Distância de Levenshtein

A transcrição foi comparada ao texto originalmente enviado ao TTS por meio da distância de Levenshtein normalizada, isto é, o número de inserções, remoções e substituições entre os dois textos em proporção ao comprimento da maior das duas cadeias, conforme a formalização apresentada na Equação 2.1 (valores de similaridade próximos de 1 indicam maior correspondência).

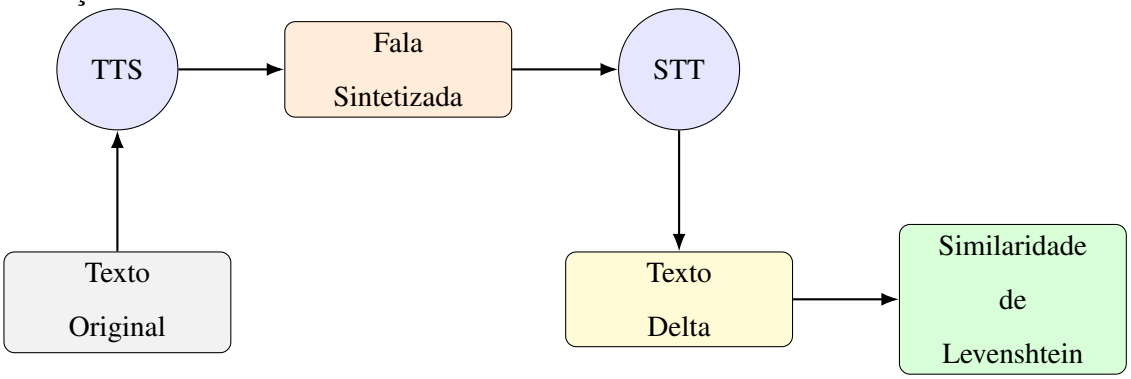
A escolha da distância de Levenshtein como indicador de inteligibilidade decorre de três fatores. O primeiro é que ela opera em nível de caractere, o que a torna sensível a erros de pronúncia parciais, em que a palavra é reconhecida de forma aproximada mas não idêntica, situação frequente em fala com sotaque e que uma métrica em nível de palavra registraria como erro total. O segundo é a normalização adotada (Marzal e Vidal, 1993), pelo comprimento da maior das duas cadeias, que limita o indicador ao intervalo $[0, 1]$ e o torna simétrico, impedindo que transcrições muito mais longas que a referência produzam valores sem limite superior. O terceiro é a comparabilidade interna: como a análise desta dissertação é relativa, isto é, compara *checkpoints* de um mesmo modelo sob o mesmo *pipeline*, uma escala limitada facilita a leitura conjunta com os preditores de MOS, também expressos em escalas limitadas.

A distância de Levenshtein relaciona-se diretamente às duas taxas de erro convencionais da área de reconhecimento automático de fala, a taxa de erro de caractere (*Character Error Rate*, CER) e a taxa de erro de palavra (*Word Error Rate*, WER) (Hunt, 1990; Morris et al., 2004): as três derivam da mesma distância de edição e diferem apenas na unidade considerada e no denominador da normalização. O CER corresponde à mesma distância em caracteres, porém dividida pelo comprimento da referência, e o WER aplica a distância sobre sequências de palavras, dividida pelo número de palavras da referência. Por serem convenção na literatura e permitirem comparação com valores publicados por outros trabalhos, CER

e WER foram derivados das mesmas transcrições e são reportados no Apêndice D, junto aos gráficos correspondentes.

O fluxo de avaliação da similaridade textual está ilustrado na Figura 4.3.

Figura 4.3: Fluxo conceitual para estimar a preservação do conteúdo linguístico por meio de transcrição automática e similaridade baseada na distância de Levenshtein.



Fonte: elaborado pelo autor (2026).

A Tabela 4.9 traz casos com distância maior (por exemplo, “Frei Damião” reconhecido como “Freio de Caminhão”). A similaridade textual, obviamente, não avalia sotaque nem prosódia.

Tabela 4.9: Exemplos de similaridade baseada na distância de Levenshtein entre o texto original e o texto reconhecido por STT.

Texto Original	**Texto Reconhecido**	**Distância de Levenshtein**	**Similaridade**
Eu te amo	Eu te amo	0	1,00
Frei Damião	Freio de Caminhão	6	0,65

Fonte: elaborada pelo autor (2026).

4.6.5 Métricas de Treinamento

Além das métricas sobre a fala sintetizada, foram consideradas métricas registradas durante o treinamento (ajuste fino e treinamento a partir de pesos aleatórios): perda por passo e por época, tempo por passo e tempo total até o critério de parada. Esses registros já orientaram o monitoramento nas Seções 4.5.2 e 4.5.3 e foram reunidos na análise da Questão de Pesquisa 3 para relacionar evolução da perda com as métricas de qualidade calculada e com a distância de Levenshtein nas mesmas condições experimentais.

4.7 Avaliação Humana do Sotaque

A avaliação humana descrita nesta seção tem por objetivo investigar a Questão de Pesquisa 4: verificar se a adaptação dos modelos produz percepção consistente do sotaque paraibano nas falas sintetizadas e se essa adaptação preserva a clareza e a naturalidade percebidas. Trata-se de uma avaliação perceptiva comparativa, complementar à avaliação automática (Seção 4.6), cujos resultados são reportados na Seção 5.5.

4.7.1 Aspectos Éticos

O instrumento de avaliação perceptiva foi submetido à apreciação do Comitê de Ética em Pesquisa (CEP) por meio da Plataforma Brasil. A pesquisa recebeu parecer consubstanciado favorável à sua realização, cujo documento na íntegra é reproduzido no Apêndice E.

4.7.2 Estrutura da Survey

A avaliação foi conduzida por meio de um formulário *on-line*, autoaplicável e de preenchimento voluntário, reproduzido na íntegra no Apêndice F. O instrumento inicia com um termo de consentimento e uma pergunta de caracterização do participante (familiaridade com o sotaque), seguidos pelos itens de avaliação propriamente ditos. Cada item de avaliação apresenta um par de áudios, identificados de forma neutra como "Áudio A" e "Áudio B", acompanhado de duas perguntas de múltipla escolha (detalhadas na Seção 4.7.4).

Os itens organizam-se em dois blocos complementares (Seção 4.7.5): o primeiro confronta cada modelo com a sua própria versão adaptada e o segundo confronta as versões adaptadas dos modelos entre si. Para reduzir o efeito de variações estocásticas pontuais de uma única síntese, cada confronto é avaliado com três amostras de áudio distintas. No total, o formulário apresenta 18 pares de áudios, cada um com duas perguntas, totalizando 36 itens de resposta. Em todos os pares, os áudios comparados correspondem à mesma sentença de texto, de modo que as diferenças percebidas decorram apenas da condição (original ou adaptada) ou do modelo, e não do conteúdo lexical.

4.7.3 Perfil dos Participantes

Os participantes foram recrutados de forma voluntária, por divulgação aberta do formulário. Para permitir análises por subgrupo, o instrumento incluiu uma pergunta de caracterização quanto à familiaridade com o sotaque paraibano, com quatro categorias mutuamente excludentes: (i) ser da Paraíba; (ii) ser familiarizado com o sotaque paraibano; (iii) ser familiarizado com sotaques nordestinos em geral; e (iv) ter pouca familiaridade com esses sotaques. Essa estratificação é relevante porque a percepção de um sotaque regional tende a ser mais acurada entre ouvintes do próprio estado, permitindo confrontar a avaliação do público geral com a do público de referência. A composição efetiva da amostra obtida é reportada na Seção 5.5.

4.7.4 Escalas de Avaliação

Para cada par de áudios, o participante respondeu a dois critérios de avaliação:

- Sotaque paraibano: "Qual áudio representa melhor o sotaque paraibano?", com as opções *Áudio A*, *Áudio B*, *Ambos aparentam ter sotaque paraibano* e *Não consigo distinguir*.
- Clareza e naturalidade: "Considerando naturalidade e clareza, qual áudio é melhor?", com as opções *Áudio A*, *Áudio B*, *Ambos soam igualmente bem* e *Não consigo distinguir*.

Optou-se por uma escala comparativa (preferência entre dois áudios), e não por uma escala absoluta do tipo MOS aplicada isoladamente a cada amostra. A comparação par a par é mais robusta para discriminar diferenças sutis, como a presença ou ausência de um sotaque regional, e reduz o viés de calibração individual, em que diferentes participantes ancoram a mesma nota em referências distintas. As opções "Ambos aparentam ter sotaque paraibano"/"Ambos soam igualmente bem" e "Não consigo distinguir" foram incluídas para capturar, respectivamente, situações de equivalência percebida (ambos os áudios apresentam a característica) e de indistinguibilidade efetiva, evitando a escolha forçada quando o participante não percebe diferença.

4.7.5 Estratégias de Comparação entre Áudios

A avaliação perceptiva restringiu-se aos três modelos que apresentaram resultados preliminares positivos nos experimentos de ajuste fino, isto é, com indícios audíveis de incorporação do sotaque paraibano na escuta de acompanhamento (Seção 5.3.2): XTTS, Fish Speech e Orpheus TTS. O F5-TTS e o Qwen3-TTS foram excluídos por não terem apresentado aproximação perceptível ao sotaque alvo; além do critério de mérito, essa restrição reduziu o custo da avaliação, evitando expor os participantes a confrontos adicionais sem expectativa de efeito. A avaliação empregou dois desenhos de comparação par a par, organizados em blocos:

- Bloco 1 (Original $\times$ Adaptado). Cada modelo (XTTS, Fish Speech e Orpheus TTS) tem a sua versão original (modelo base) confrontada com a sua versão adaptada por ajuste fino, em três amostras. Este bloco responde diretamente à Questão de Pesquisa 4: se o ajuste fino adiciona percepção de sotaque e se preserva a clareza e a naturalidade.

- Bloco 2 (Ranqueamento entre os modelos adaptados). As versões adaptadas dos três modelos são confrontadas entre si, em todos os pares possíveis (XTTS $\times$ Fish Speech, Fish Speech $\times$ Orpheus TTS e XTTS $\times$ Orpheus TTS), também com três amostras por confronto. Este bloco permite ordenar os modelos quanto à qualidade do sotaque e à clareza/naturalidade percebidas.

A escolha pelo desenho par a par, em vez de avaliação isolada de cada áudio, decorre dos mesmos motivos apontados na Seção 4.7.4: comparações diretas tornam a tarefa do ouvinte mais simples e mais sensível a diferenças de pequena magnitude.

4.7.6 Randomização das Amostras

A posição dos áudios em cada item (qual condição ou modelo aparece como “Áudio A” e qual aparece como “Áudio B”) foi aleatorizada por questão. Dessa forma, o rótulo “Áudio A” não corresponde a um modelo ou condição fixos ao longo do formulário, o que neutraliza eventuais vieses de posição (por exemplo, tendência a preferir sistematicamente a primeira ou a

segunda opção). O mapeamento entre os rótulos A/B e as respectivas condições ou modelos foi mantido em um gabarito separado, utilizado apenas na etapa de análise (Seção 4.7.9).

4.7.7 Controle de Viés

Além da randomização de posição, adotaram-se outras medidas para reduzir vieses de resposta. Em nenhum momento os modelos foram identificados aos participantes: os áudios foram apresentados de forma anonimizada, sem menção a nome de arquitetura, de fabricante ou à condição (original/adaptado), evitando que o conhecimento prévio sobre um sistema influenciasse o julgamento. A comparação foi feita sempre sobre a mesma sentença, isolando a variável de interesse. A inclusão das opções "Ambos aparentam ter sotaque paraibano"/"Ambos soam igualmente bem" e "Não consigo distinguir" mitiga o viés de escolha forçada, permitindo que o participante expresse equivalência ou ausência de percepção de diferença em vez de adivinhar. Por fim, o consentimento informado e a apresentação puramente sonora (sem exibição de texto que revelasse a origem dos áudios) reforçam a neutralidade do instrumento.

4.7.8 Controle de Fadiga

A fadiga do avaliador é uma ameaça reconhecida em testes perceptivos longos, pois reduz a atenção e a consistência das respostas ao final do instrumento. Para mitigá-la, o formulário foi mantido enxuto: 18 pares de áudios, com amostras curtas e tarefa de resposta simples (múltipla escolha entre quatro opções), em uma única sessão de duração reduzida. O tempo médio de execução do formulário foi de aproximadamente dez minutos.

4.7.9 Critérios de Análise dos Resultados

As respostas de cada modelo foram agregadas sobre as três amostras de áudio correspondentes. Como o objetivo do trabalho é verificar se a fala adaptada carrega o sotaque paraibano, e não apenas vencer o original em uma disputa direta, adotaram-se as seguintes convenções de positividade na análise:

- na percepção de sotaque, considera-se favorável a resposta que escolhe o áudio adaptado ou a opção "ambos aparentam ter sotaque paraibano";

- na clareza e naturalidade, considera-se que "não houve piora" quando a resposta escolhe o áudio adaptado ou a opção "ambos soam igualmente bem";
- a resposta "não consigo distinguir" é tratada como neutra/inconclusiva e excluída dos testes estatísticos.

Para o Bloco 1 (Original $\times$ Adaptado), a significância foi avaliada por teste binomial exato (adaptado *versus* original, descartados os empates e os casos indistinguíveis; hipótese nula de proporção igual a $0{,}5$) (Agresti, 2013). A escolha decorre da estrutura dos dados: excluídas as respostas neutras, cada resposta é um ensaio de Bernoulli com dois desfechos possíveis e, sob a hipótese de indistinguibilidade entre os áudios, probabilidade $0{,}5$ para cada um, de modo que a contagem de respostas favoráveis segue exatamente uma distribuição binomial (Seção 2.7.5). Para as comparações de proporções entre modelos distintos (grupos independentes), emprega-se o teste exato de Fisher, com correção de Holm para múltiplas comparações (Agresti, 2013; Holm, 1979). Para o Bloco 2 (ranqueamento entre os modelos adaptados), reporta-se a taxa de vitória de cada modelo, definida como a razão entre vitórias e o total de confrontos decididos (vitórias mais derrotas, excluídos os empates). Todas as análises foram recalculadas em três recortes da amostra (todos os participantes; apenas os participantes da Paraíba; e o conjunto de participantes familiarizados, excluindo a categoria de pouca familiaridade), de modo a verificar a robustez das conclusões em relação ao perfil do ouvinte. Como o teste binomial por item agrega três respostas de um mesmo participante (três amostras por confronto), essas respostas não são totalmente independentes, havendo dependência intra-avaliador; por isso, além dos recortes por subgrupo, conduziu-se também uma análise conservadora por participante, em que cada respondente contribui com um único voto por modelo (a escolha majoritária entre as três amostras), reduzindo o tamanho efetivo de 168 respostas para 56 participantes.

4.8 Considerações Éticas

A pesquisa envolveu dados de fala humana em duas frentes, e em ambas adotaram-se cuidados éticos. Na construção do corpus *FALA_PB*, as bases públicas reaproveitadas foram utilizadas conforme suas licenças de uso, e a coleta colaborativa contou com a participação

voluntária e informada dos colaboradores, que cederam suas gravações cientes da finalidade de pesquisa. Não foram coletados dados pessoais sensíveis, e os áudios foram tratados de forma a preservar a identidade dos falantes.

Na avaliação perceptiva, a participação também foi voluntária e precedida de termo de consentimento. O formulário não coletou dados que permitissem identificar os respondentes, registrando apenas a familiaridade com o sotaque, necessária às análises por subgrupo. Os modelos avaliados foram apresentados de forma anonimizada, sem identificação de arquitetura, fabricante ou criador, garantindo a imparcialidade do julgamento e evitando qualquer associação a interesses comerciais.

4.9 Limitações Metodológicas

Algumas limitações metodológicas devem ser registradas. O corpus *FALA_PB*, embora adequado aos experimentos, é fortemente dominado pelo CoLingPB e não representa estatisticamente toda a variabilidade do sotaque paraibano, o que delimita o alcance das generalizações. Os experimentos foram conduzidos sob restrição real de infraestrutura, com uma única estação de trabalho com GPU, o que limitou a escala dos treinos e tornou inviável o treinamento completo (a partir de pesos aleatórios iniciais) em larga escala, conduzido apenas como estudo de caso em um modelo.

Na avaliação automática, os preditores de MOS utilizados são estimativas relativas, treinadas majoritariamente em fala neutra e, por isso, sensíveis a características de sotaque, mesmo na ausência de perda real de qualidade. Por esse motivo, devem ser interpretados de forma comparativa e em conjunto, e não como substitutos do MOS subjetivo. Na avaliação perceptiva, o número de participantes é modesto (56 respondentes) e a amostra foi obtida por adesão voluntária, sem controle rígido de representatividade demográfica.

4.10 Síntese do Capítulo

Este capítulo apresentou a metodologia da pesquisa. Foram definidas as questões de pesquisa e descritos os materiais e modelos utilizados, a construção do corpus *FALA_PB* a partir da integração de bases públicas e de uma coleta colaborativa, e as três estratégias de adaptação

investigadas: clonagem de fala, ajuste fino e treinamento a partir de pesos aleatórios. Em seguida, detalharam-se os protocolos de avaliação automática (preditores de MOS e similaridade de Levenshtein) e de avaliação perceptiva com ouvintes, além das considerações éticas e das limitações metodológicas. Os resultados obtidos com esses procedimentos são apresentados no Capítulo 5.

Capítulo 5

Experimentos e Resultados

5.1 Visão Geral dos Experimentos

Esta seção apresenta uma visão geral dos experimentos realizados nesta pesquisa, incluindo a construção do corpus FALA_PB, os experimentos de adaptação dos modelos de síntese de fala e as etapas de avaliação automática e manual utilizadas para análise dos resultados obtidos.

Os experimentos foram organizados em torno das quatro questões de pesquisa da dissertação, apresentadas no Capítulo 1, investigando a construção de uma base de fala adequada ao sotaque paraibano, a capacidade de diferentes estratégias incorporarem o sotaque regional, os impactos das adaptações na qualidade e na inteligibilidade da fala sintetizada e a avaliação humana do sotaque nas falas geradas.

5.2 Caracterização do Corpus FALA_PB

Esta seção apresenta os resultados obtidos na construção e caracterização do corpus *FALA_PB*. Os procedimentos metodológicos correspondentes estão descritos no capítulo anterior, Seção 4.4 e, quanto às fontes utilizadas, Seção 4.3.1.

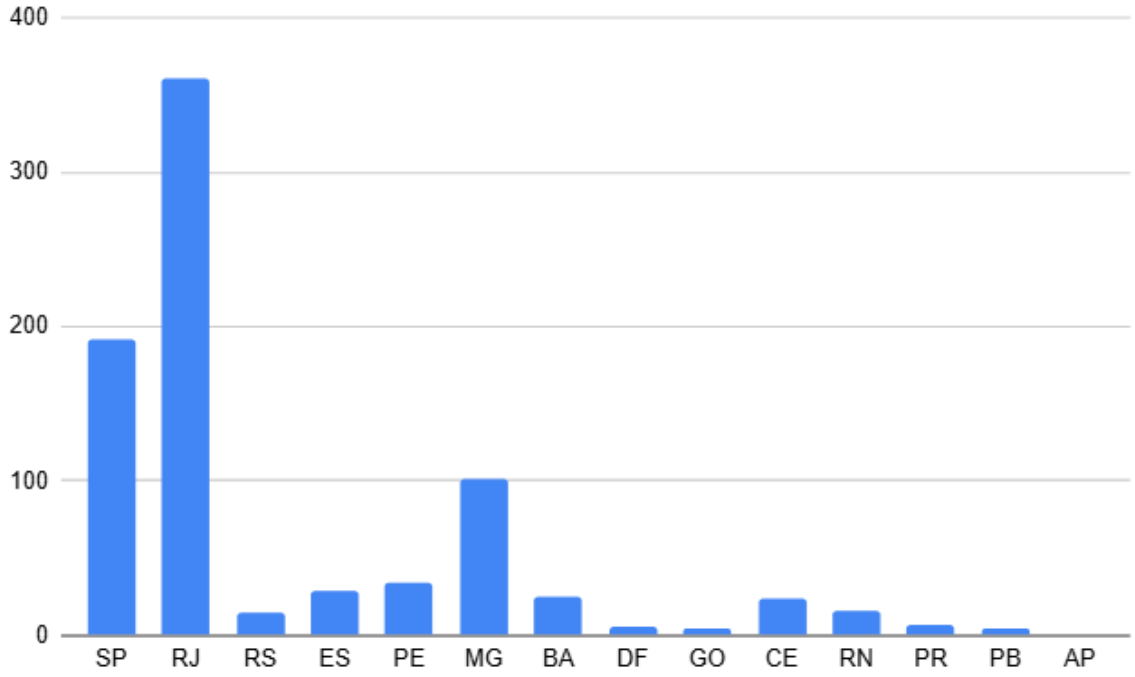
5.2.1 Distribuição das Bases Integradas

Sotaque Brasileiro e Braccent

Durante o levantamento e a filtragem das bases Sotaque Brasileiro (Milan, 2021) e Braccent (Batista, 2019) para o sotaque paraibano, identificou-se desbalanceamento regional na distribuição de falantes e volume insuficiente de amostras nordestinas para uso isolado nos experimentos.

Na base Sotaque Brasileiro, observou-se maior concentração de amostras provenientes de estados como Rio de Janeiro e São Paulo, enquanto estados nordestinos apresentam quantidade reduzida de dados, conforme a Figura 5.1.

Figura 5.1: Histograma de distribuição por estado dos falantes do *dataset* Sotaque Brasileiro.

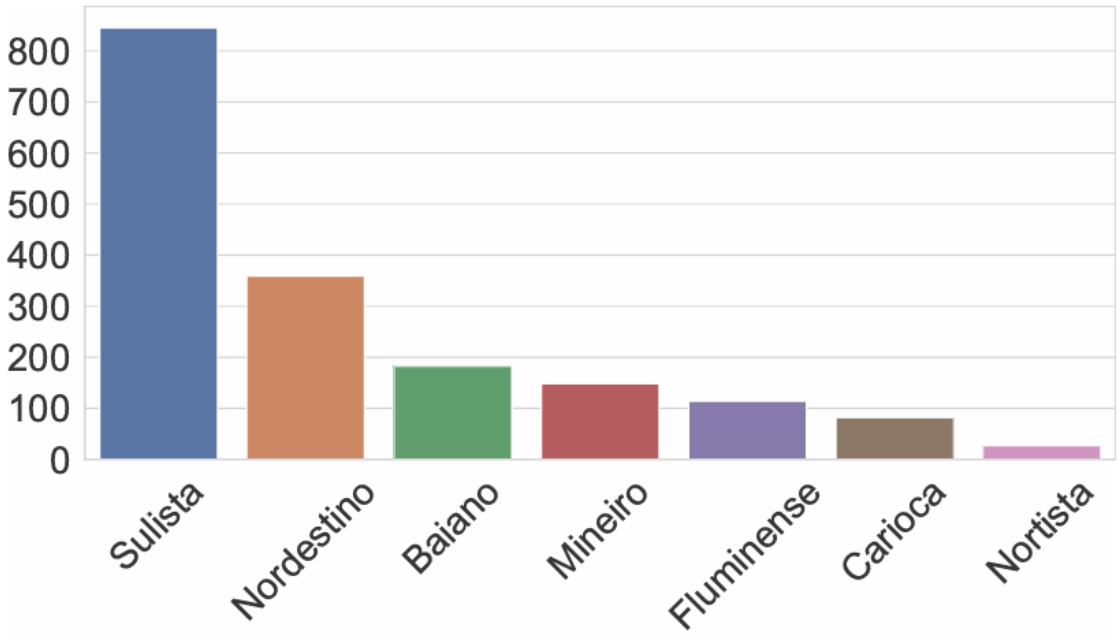


Fonte: Milan (2021).

Para integração ao *FALA_PB*, foram selecionadas apenas quatro falas associadas ao estado da Paraíba.

Na base Braccent, identificou-se predominância de sotaques das regiões Sul e Sudeste, com quantidade mais reduzida de amostras nordestinas, conforme a Figura 5.2.

Figura 5.2: Histograma de distribuição por sotaque dos falantes do *dataset* Braccent.



Fonte: Batista (2019).

A identificação das amostras paraibanas não se apoiou no rótulo de sotaque exibido na Figura 5.2, que agrega as gravações em categorias regionais mais amplas. A base é distribuída com um arquivo de metadados no formato *.csv* (*Comma-Separated Values*) que registra, para cada áudio, o estado e o município de origem do falante; a seleção foi feita filtrando esse campo pela Paraíba. Por esse critério, foram incorporadas ao corpus aproximadamente 64 amostras relacionadas ao sotaque paraibano.

Corpus Linguístico da Paraíba (CoLingPB)

O CoLingPB (Stein et al., 2013) reúne 324 entrevistas coletadas em diferentes regiões do estado da Paraíba, identificadas pelo esquema de codificação original descrito no Apêndice B (Tabela B.1). A Tabela 5.1 apresenta a distribuição dessas entrevistas por município.

Tabela 5.1: Distribuição de entrevistas do CoLingPB por município.

Município	**Quantidade de Entrevistas**
João Pessoa	16
Conceição	15
Itaporanga	14
Cajazeiras	13
Princesa Isabel	13
Catingueira	13
Belém	12
Guarabira	12
Itabaiana	12
Esperança	12
Campina Grande	12
Umbuzeiro	12
Cuité	12
Barra de Santa Rosa	12
Picuí	12
Soledade	12
Cabaceiras	12
Serra Branca	12
Congo	12
Taperoá	12
Monteiro	12
Patos	12
Catolé do Rocha	12
Pombal	12
Sousa	12
Mamanguape	12
Total	**324**

Fonte: elaborada pelo autor (2026), com dados de (Stein et al., 2013).

A Tabela 5.1 refere-se às entrevistas originais. A distribuição dos segmentos após filtragem, com metadados recuperados a partir da codificação das entrevistas, é caracterizada em detalhe no Apêndice B (Seção B.4), por zona, sexo, faixa etária, duração e número de palavras por segmento.

5.2.2 Resultados das Etapas de Filtragem e Segmentação

A segmentação automática do CoLingPB utilizou as marcações de tempo e interlocutor dos arquivos *.trs*[1], com cortes executados por meio da ferramenta FFmpeg. Após segmentação e filtragem, o conjunto compreende 33.649 arquivos com extensão *.wav*[2], correspondentes à fala dos entrevistados, com duração total de 108 h 5 min 39 s. A Tabela 5.2 reúne essas métricas, consistentes com os totais consolidados do corpus na Tabela 5.4.

Cada segmento recebeu um nome que preserva a entrevista de origem e o intervalo do corte, segundo o padrão de nomenclatura detalhado no Apêndice B.

Tabela 5.2: Métricas do CoLingPB após extração, segmentação e filtragem.

Métrica	**Valor**
Segmentos	33.649
Duração total dos cortes	108 h 5 min 39 s
Duração média dos segmentos	11,56 s

Fonte: elaborada pelo autor (2026), com dados de (Stein et al., 2013).

Durante a escuta e validação manual dos segmentos, constatou-se que parte dos áudios das entrevistas originais apresentava ruídos e sons ambientes inadequados para treinamento de modelos de síntese de fala, entre eles ruídos de veículos, sons de pássaros, reverberação, conversas paralelas e demais interferências. A filtragem automática foi aplicada sobre o conjunto segmentado com base no preditor *DNSMOS* (Reddy et al., 2022), adotando-se como critério de exclusão o escore *ovrl* inferior a 3, limiar definido manualmente a partir da escuta de amostras. O volume de 33.649 segmentos reportado acima corresponde ao conjunto já segmentado e filtrado.

[1] A extensão *.trs* é usada para armazenar uma transcrição de áudio estruturada em formato XML, contendo o texto falado junto com o tempo exato de cada fala, o nome de quem fala e os blocos de áudio.

[2] A extensão *.wav* é usada para guardar áudio sem compressão, na forma de um vetor de amplitudes.

5.2.3 Resultados da Coleta Colaborativa

Após filtragem e organização dos arquivos, a coleta colaborativa reuniu 1.149 amostras de fala, totalizando 3 h 25 min 39 s, conforme a Tabela 5.3. O subconjunto via WhatsApp contribuiu com 93 gravações, enviadas pelos participantes a partir dos próprios celulares. O subconjunto de leitura de *Vidas Secas* concentrou 1.056 trechos (3 h 8 min 6 s), obtidos com o procedimento de gravação contínua e segmentação automática nas pausas descrito na metodologia, em sala silenciosa com microfone Arcano Tsion.

Tabela 5.3: Composição final das amostras da coleta colaborativa.

Subconjunto	**Amostras**	**Duração total**
Coleta colaborativa WhatsApp	93	0 h 17 min 33 s
Coleta colaborativa com leitura de *Vidas Secas*	1.056	3 h 8 min 6 s
Total	**1.149**	**3 h 25 min 39 s**

Fonte: elaborada pelo autor (2026).

As etapas de filtragem, padronização e controle de qualidade aplicadas a esses subconjuntos seguiram os mesmos critérios das demais fontes, resultando na versão consolidada do corpus utilizada nos experimentos desta dissertação.

5.2.4 Composição Final do Corpus

A Tabela 5.4 resume a composição quantitativa das fontes integradas ao *FALA_PB* após ajustes, filtros e padronização final (áudio mono, 24 kHz, PCM 16 bits). O CoLingPB constitui a principal fonte de volume do corpus, complementado pela coleta colaborativa e por um pequeno conjunto de amostras das bases Sotaque Brasileiro e Braccent, totalizando aproximadamente 111 horas de fala. A coleta colaborativa agrega os subconjuntos obtidos por WhatsApp (93 amostras) e a leitura de trechos de *Vidas Secas* (1.056 amostras), detalhados na Tabela 5.3. As métricas intermediárias do CoLingPB, ainda a 16 kHz, constam da Tabela 5.2.

Após integração das fontes, o corpus *FALA_PB* reuniu amostras da Sotaque Brasileiro, Braccent, CoLingPB e da coleta colaborativa, com metadados textuais organizados conforme a origem de cada fonte. O corpus constitui a principal base dos experimentos e encontra-se

Tabela 5.4: Resumo das bases de dados de fala utilizadas na construção da base FALA_PB.

Base de Dados	**Quantidade de Áudios**	**Duração Total**
Sotaque Brasileiro	4	0 h 0 min 53 s
Braccent	64	0 h 13 min 34 s
CoLingPB (processado)	33.649	108 h 5 min 39 s
Coleta colaborativa	1.149	3 h 25 min 39 s
Total	**34.866**	**111 h 45 min 45 s**

Fonte: elaborada pelo autor (2026), com dados de (Milan, 2021; Batista, 2019; Stein et al., 2013).

disponibilizado publicamente na plataforma Zenodo (de Morais, 2026)[3].

5.2.5 Limitações Identificadas no Corpus

Entre as principais limitações observadas destacam-se:

1. volume reduzido de falas paraibanas nas bases Sotaque Brasileiro e Braccent;
2. desbalanceamento regional nas bases públicas analisadas;
3. heterogeneidade de qualidade, especialmente em trechos do CoLingPB;
4. impossibilidade de caracterizar de forma ampla a diversidade de falantes no subconjunto colaborativo, em função da opção por não divulgar contagens nem identificadores pessoais.

5.2.6 Discussão sobre a Questão de Pesquisa 1

A Questão de Pesquisa 1 indaga como construir uma base de fala do sotaque paraibano adequada ao treino e à adaptação de modelos de síntese. A resposta apoiada nesta dissertação é a constituição do corpus *FALA_PB* por integração e processamento de fontes complementares, e não por uma única base pronta:

- incorporação de amostras paraibanas das bases Sotaque Brasileiro e Braccent (volume pequeno, papel complementar);

[3] `https://zenodo.org/records/21251709`

- processamento do CoLingPB (extração de segmentos a partir de entrevistas com arquivos *.trs*, segmentação e filtragem acústica), principal fonte em horas de fala;
- coleta colaborativa em dois subconjuntos: três frases por participante via WhatsApp (93 amostras) e leitura de *Vidas Secas* com segmentação automática nas pausas (1.056 amostras);
- filtragem e controle de qualidade, com remoção de trechos ruidosos ou de baixa inteligibilidade;
- padronização final de todos os áudios para o formato *.wav* mono, com codificação PCM em 16 bits e amostragem a 24 kHz; os segmentos do CoLingPB, gerados a 16 kHz, foram reamostrados nesta etapa.

O conjunto obtido é adequado aos experimentos realizados neste trabalho (Tabela 5.4), embora não constitua uma amostra estatisticamente representativa de toda a variabilidade do sotaque paraibano, pelas limitações já apresentadas. Nesse sentido, o *FALA_PB* responde à Questão de Pesquisa 1 ao oferecer um corpus consolidado, documentado e reutilizável para síntese com sotaque regional. O quadro seguinte resume a resposta dada à Questão de Pesquisa 1.

Questão de Pesquisa 1: Como construir uma base de dados de fala voltada ao sotaque paraibano adequada para treinamento e adaptação de modelos de síntese de fala?
Resposta: Pela integração e processamento de fontes complementares, e não por uma única base pronta. O corpus *FALA_PB* foi constituído pela incorporação de amostras paraibanas das bases Sotaque Brasileiro e Braccent, pelo processamento e segmentação do CoLingPB (principal fonte em horas de fala) e por uma coleta colaborativa complementar com falantes da Paraíba, seguidos de filtragem acústica e padronização final (*.wav* mono, 24 kHz, PCM 16 bits). O conjunto resultante (Tabela 5.4) mostrou-se adequado aos experimentos de adaptação realizados nesta dissertação, embora não constitua amostra estatisticamente representativa de toda a variabilidade do sotaque paraibano, limitação documentada nesta seção.

5.3 Resultados das Estratégias de Adaptação

5.3.1 Resultados da Clonagem de Fala

Esta seção apresenta os resultados dos experimentos exploratórios de clonagem de fala descritos na Seção 4.5.1. O procedimento utilizou amostras de referência do corpus *FALA_PB* (10, 30 e 60 s) e síntese de textos distintos do conteúdo das referências.

Preservação de Timbre e Sotaque

Os modelos reproduziram de forma satisfatória o timbre e a identidade vocal do locutor de referência. O sotaque paraibano, entretanto, não foi capturado em nenhuma das condições testadas: mesmo quando o texto sintetizado era parecido com o conteúdo do áudio usado na clonagem, a fala gerada preservou apenas as características de timbre, sem traços perceptíveis da pronúncia regional.[4]

Limitações Observadas

Independentemente da duração da referência (10, 30 ou 60 s) e da proximidade entre o texto de síntese e o conteúdo clonado, o resultado se manteve o mesmo: a voz gerada soava como o locutor de referência em timbre, mas com a pronúncia do modelo base, sem incorporação do sotaque paraibano. Além disso, quando o texto de síntese se afastava do conteúdo utilizado na clonagem, observou-se degradação adicional de qualidade e naturalidade.

Discussão: Por que a Clonagem Não Incorporou o Sotaque?

O resultado obtido contrasta, à primeira vista, com a literatura que relata sucesso da clonagem *zero-shot*. Esse contraste, contudo, é menos direto do que parece, uma vez que os trabalhos em questão não relatam necessariamente o mesmo sucesso que se buscava aqui. Modelos como o XTTS (Casanova et al., 2024) reportam alta fidelidade na reprodução da identidade vocal a partir de poucos segundos de referência, e é exatamente esse comportamento que se observou nos experimentos desta seção: o timbre do locutor paraibano foi

[4] Amostras de áudio das três durações de referência disponíveis em `https://thomazdiniz.github.io/fala_pb/amostras/`.

reproduzido de forma satisfatória. O que não se transferiu foi o sotaque, isto é, um conjunto de traços de pronúncia e de prosódia que os trabalhos de clonagem, em geral, não se propõem a avaliar. Por outro lado, os trabalhos que relatam transferência bem-sucedida de sotaque recorrem a arquiteturas que tratam o sotaque como fator explícito e separável da identidade do falante (Ju et al., 2024; Annamdevula et al., 2026), a procedimentos dedicados de conversão de sotaque (Liu et al., 2020) ou a adaptação com dados da variedade alvo (Zhou et al., 2023), e não à clonagem *zero-shot* com pesos congelados.

Essas abordagens não foram investigadas nesta dissertação por três razões. A primeira é de escopo: a pesquisa delimitou-se à adaptação de modelos abertos já pré-treinados, e não à concepção de arquiteturas que modelem o sotaque como fator explícito. A segunda é de dados: representações desacopladas de locutor e de sotaque são aprendidas a partir de corpora multi-sotaque rotulados por variedade, condição que não se verifica para o português brasileiro, conforme discutido na Seção 3.3.3. Das bases regionais disponíveis foi possível aproveitar pouco mais de quatorze minutos de fala paraibana. A terceira é de infraestrutura: tais arquiteturas exigem treinamento a partir de pesos aleatórios ou adaptação extensiva, estratégia cuja inviabilidade nas condições desta pesquisa foi verificada experimentalmente na Seção 5.3.3. Registre-se ainda que o trabalho de Annamdevula et al. (2026) é contemporâneo à execução dos experimentos aqui relatados. Essa direção é retomada como sugestão de trabalho futuro no Capítulo 6.

A partir dessa leitura, formulam-se duas hipóteses, não excludentes, para o resultado observado. A primeira é que os modelos empregados não foram pré-treinados com variedade suficiente de sotaques para que o sotaque paraibano estivesse representado em seu espaço de saída. Como a clonagem *zero-shot* não altera os pesos, ela apenas seleciona e recombina aquilo que o modelo já aprendeu a produzir durante o pré-treinamento; se a variedade regional alvo não figura, ou figura de forma marginal, nos dados de pré-treinamento, nenhuma referência fornecida na inferência será capaz de fazer o modelo produzi-la.

A segunda hipótese é que a representação extraída do áudio de referência codifica essencialmente o timbre, e não os elementos constitutivos do sotaque. Conforme discutido na Seção 2.6.3, o codificador de voz produz uma representação global da identidade vocal, otimizada para distinguir falantes entre si. A pronúncia dos fonemas e os padrões prosódicos, por sua vez, continuam determinados pelos pesos congelados do modelo, que não são

afetados pela referência. Sob essa hipótese, o comportamento observado não é uma falha da clonagem, mas o funcionamento esperado de um mecanismo cujo objetivo é justamente separar quem fala daquilo que é dito e de como é pronunciado. A verificação empírica de qualquer uma das duas hipóteses exige a caracterização acústico-fonética proposta como trabalho futuro no Capítulo 6, que permitiria medir quais dimensões da fala de referência são efetivamente preservadas na fala sintetizada.

Por isso, a clonagem mostrou-se útil apenas para imitar a voz do falante, e não para incorporar o sotaque paraibano. Optou-se, então, por não usá-la como estratégia principal, priorizando nas etapas seguintes o ajuste fino e o treinamento com o corpus *FALA_PB*.

5.3.2 Resultados do Ajuste Fino de Modelos Pré-Treinados

Esta seção apresenta os resultados dos ajustes finos descritos na Seção 4.5.2, com métricas de treinamento, curvas de perda e avaliação qualitativa preliminar.

XTTS

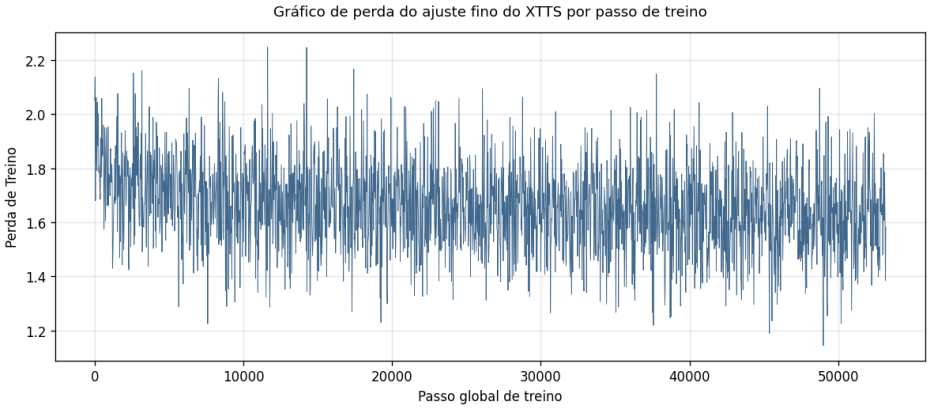
O treinamento do modelo XTTS durou 4 h 27 min em três épocas (Tabela 5.5). A perda por passo foi ruidosa, com leve tendência de queda (Figura 5.3); por época, a redução ficou mais visível (Figura 5.4).

Tabela 5.5: Resumo do treinamento do modelo XTTS.

Métrica	**Valor**
Épocas executadas	3
Tempo total de treinamento	4 horas e 27 minutos
Tempo médio por passo	0,1288 segundos

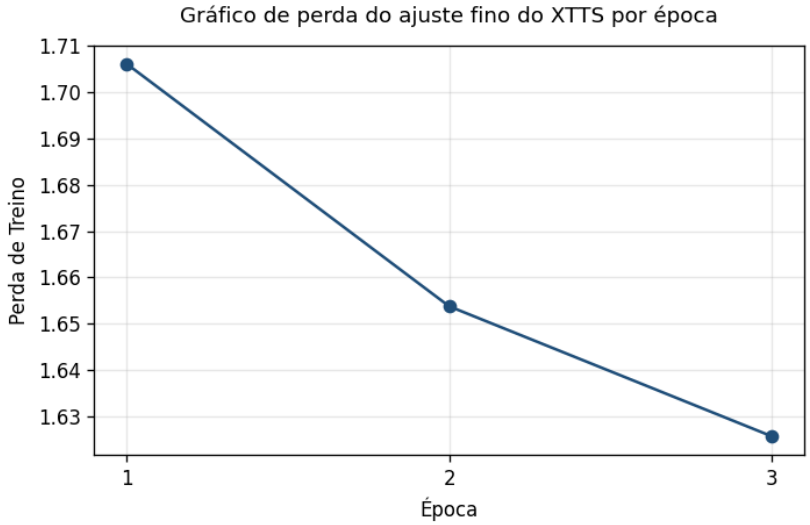
Fonte: elaborada pelo autor (2026).

Figura 5.3: Curva de perda por passo no ajuste fino do XTTS.



Fonte: elaborado pelo autor (2026).

Figura 5.4: Curva de perda por época de treinamento no ajuste fino do XTTS.



Fonte: elaborado pelo autor (2026).

Nas cinco frases de acompanhamento, a síntese mostrou alteração perceptível em relação ao modelo pré-treinado, com sotaque ainda misto entre o padrão do *checkpoint* (forte presença de falantes do Sudeste) e traços paraibanos. O ajuste incorporou o sotaque alvo apenas de forma parcial; volumes maiores de dados regionais podem ampliar a adaptação em trabalhos futuros.[5]

[5]Amostras de áudio do XTTS, nas condições original e ajustada, disponíveis em `https://thomazdiniz.github.io/fala_pb/amostras/`.

F5-TTS

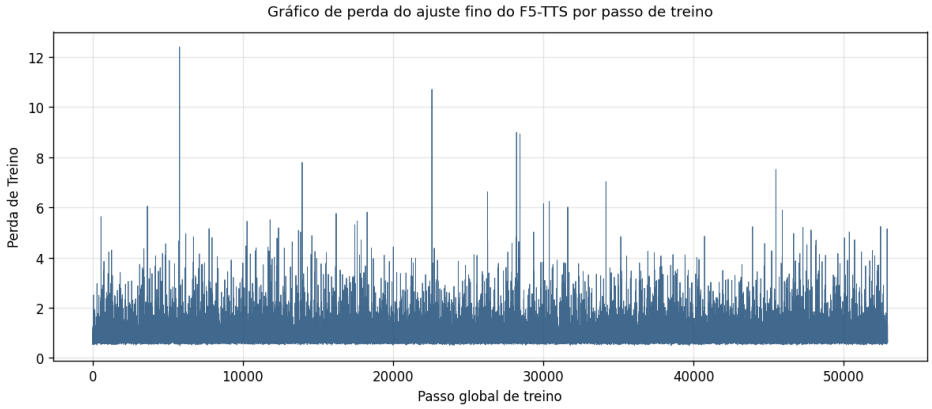
O treinamento do modelo F5-TTS durou 5 h 13 min em três épocas (Tabela 5.6). As curvas de perda (Figuras 5.5 e 5.6) foram mais ruidosas que no XTTS.

Tabela 5.6: Resumo do treinamento do F5-TTS.

Métrica	**Valor**
Épocas executadas	3
Tempo total de treinamento	5 horas e 13 minutos
Tempo médio por passo	0,342 segundos

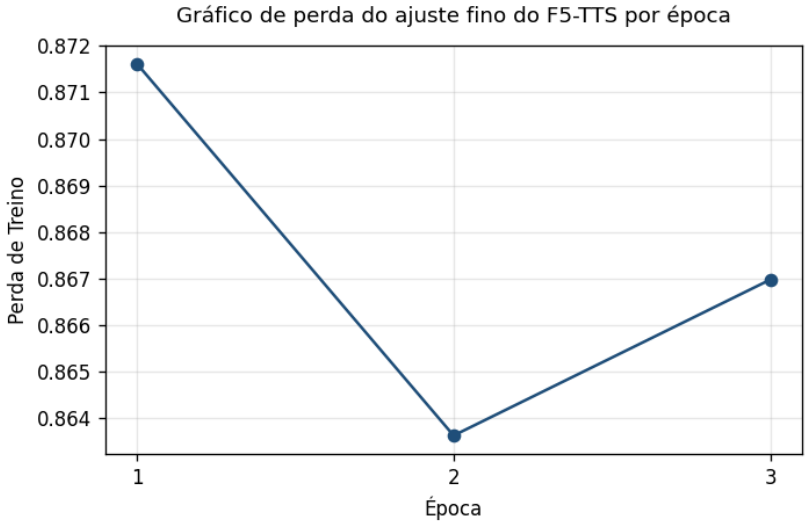
Fonte: elaborada pelo autor (2026).

Figura 5.5: Curva de perda por passo no ajuste fino do F5-TTS.



Fonte: elaborado pelo autor (2026).

Figura 5.6: Curva de perda por época de treinamento no ajuste fino do F5-TTS.



Fonte: elaborado pelo autor (2026).

Os arquivos de pesos mudaram ao final do treinamento, mas nas verificações com sementes de aleatoriedade fixas os áudios do *checkpoint* original e dos ajustados permaneceram praticamente iguais (mesma entonação e estrutura) e até mesmo sinal de áudio e formas de onda.[6]

Uma hipótese é que o ajuste tenha preservado sobretudo a capacidade geral de síntese, sem consolidar traços regionais. Alterações pontuais poderiam ser revertidas em passos seguintes. Diante disso, foram planejados experimentos de treinamento com pesos iniciais aleatórios (Seção 4.5.3).

Orpheus TTS

O treinamento do modelo Orpheus TTS, em apenas uma época, levou 21 h 22 min (Tabela 5.7). A perda por passo está na Figura 5.7.

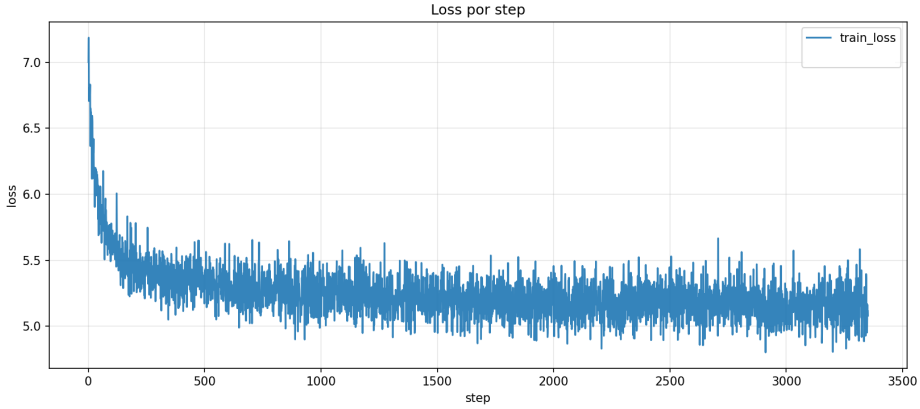
[6]Amostras de áudio do F5-TTS, nas condições original e ajustada, disponíveis em `https://thomazdiniz.github.io/fala_pb/amostras/`.

Tabela 5.7: Resumo do treinamento do Orpheus TTS.

Métrica	**Valor**
Quantidade de épocas	1
Tempo total de treinamento	21 horas e 22 minutos
Tempo médio por passo	22,3 segundos

Fonte: elaborada pelo autor (2026).

Figura 5.7: Curva de perda por passo no ajuste fino do Orpheus TTS.



Fonte: elaborado pelo autor (2026).

Nas frases de validação, o modelo mostrou incorporação satisfatória do sotaque paraibano.[7] O *checkpoint* base já apresentava alguns artefatos na síntese; mesmo assim, os resultados foram considerados relevantes para seguir nas avaliações perceptivas.

Fish Speech

O treinamento do modelo Fish Speech, em apenas uma época, levou 44 h 43 min (Tabela 5.8). A perda por passo (Figura 5.8) foi ruidosa, mas caiu entre o início e o fim do treinamento.

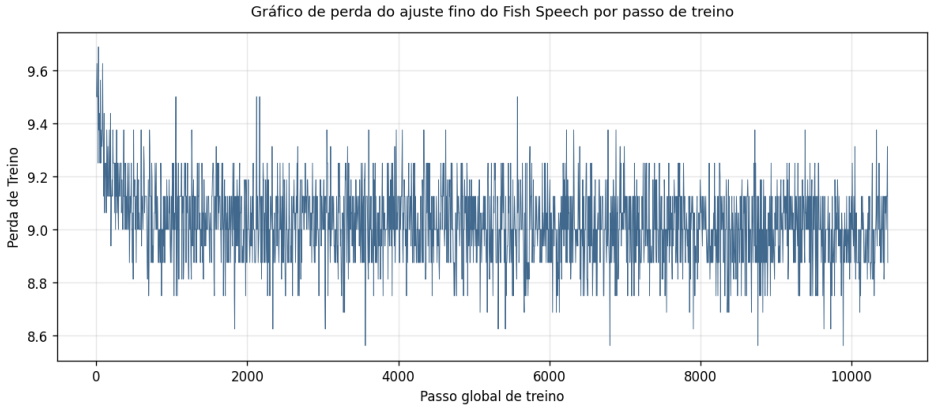
Tabela 5.8: Resumo do treinamento do Fish Speech.

Métrica	**Valor**
Quantidade de épocas	1
Tempo total de treinamento	44 horas e 43 minutos

Fonte: elaborada pelo autor (2026).

[7] Amostras de áudio do Orpheus TTS, nas condições original e ajustada, disponíveis em `https://thomazdiniz.github.io/fala_pb/amostras/`.

Figura 5.8: Curva de perda por passo no ajuste fino do Fish Speech.



Fonte: elaborado pelo autor (2026).

Nas frases de validação, em parte das amostras o sotaque paraibano ficou perceptível já após uma época.[8]

Qwen3-TTS

O treinamento do modelo Qwen3-TTS, em apenas uma época, levou 4 dias, 3 horas e 28 minutos, com 20.976 passos (Tabela 5.9). A curva de perda de treino (Figura 5.9) mostra queda ao longo de dezenas de milhares de passos, coerente com o custo computacional elevado deste experimento.

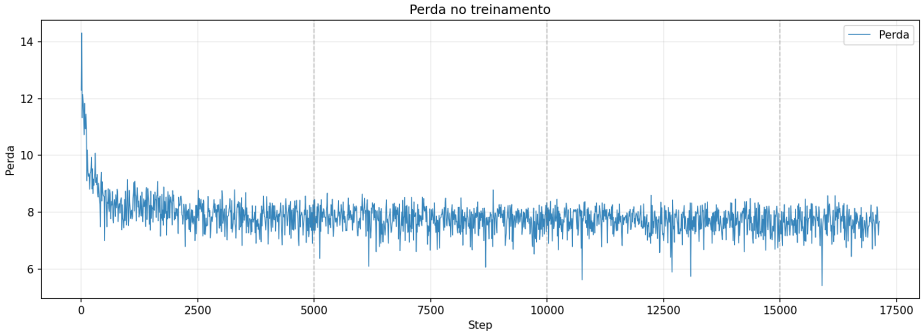
Tabela 5.9: Resumo do treinamento do Qwen3-TTS.

Métrica	**Valor**
Épocas executadas	1
Tempo total de treinamento	4 dias, 3 horas e 28 minutos

Fonte: elaborada pelo autor (2026).

[8] Amostras de áudio do Fish Speech, nas condições original e ajustada, disponíveis em `https://thomazdiniz.github.io/fala_pb/amostras/`.

Figura 5.9: Curva de perda de treino (*train loss*) no ajuste fino do Qwen3-TTS.



Fonte: elaborado pelo autor (2026).

Apesar da redução da perda de treino ao longo de dezenas de milhares de passos e de mais de três dias de computação, as sínteses nos checkpoints avaliados não evidenciaram aproximação audível ao sotaque paraibano[9] em relação ao modelo base. Por isso, esta configuração do Qwen3-TTS foi excluída da avaliação perceptiva com participantes (Seção 5.5), permanecendo documentada aqui pelos custos de treino e pelo contraste entre métrica de otimização e percepção do sotaque.

5.3.3 Resultados do Treinamento a partir de Pesos Aleatórios

Esta seção apresenta os resultados do treinamento com pesos aleatórios descrito na Seção 4.5.3. A etapa foi realizada apenas com o F5-TTS (Seção 4.5.3).

F5-TTS

Entre os modelos da Seção 4.3.2, somente o F5-TTS foi treinado a partir de pesos aleatórios, como estudo de caso da estratégia mais custosa da Questão de Pesquisa 2. A escolha repetiu os critérios da metodologia: ajuste fino sem ganho audível de sotaque e fluxo de código mais direto que o dos demais repositórios analisados.

O treinamento deste modelo partiu de pesos aleatórios, sem *checkpoint* pré-treinado nem ajuste fino prévio, sobre o mesmo corpus das rodadas anteriores de ajuste fino (Tabela 5.4). A configuração e os principais indicadores de execução estão resumidos na Tabela 5.10: a execução levou aproximadamente dez dias e a perda de validação caiu ao longo das épocas

[9] Amostras de áudio disponíveis em `https://thomazdiniz.github.io/fala_pb/amostras/`.

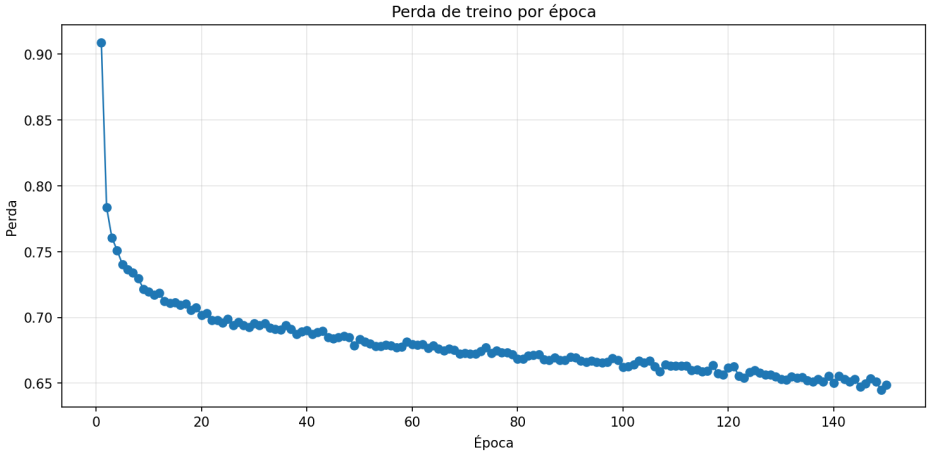
(Figura 5.10), sem que a síntese se tornasse utilizável.

Tabela 5.10: Resumo da execução do treinamento do F5-TTS a partir de pesos aleatórios.

Métrica	**Valor**
Épocas executadas	150
Tempo de execução	≈244 h (aproximadamente 10 dias)
Menor perda de validação	0,590 (época 136)
Perda de validação final	0,644 (época 150)
Checkpoint avaliado	Final (época 150)

Fonte: elaborada pelo autor (2026).

Figura 5.10: Perda de validação por época no treinamento do F5-TTS a partir de pesos aleatórios (150 épocas).



Fonte: elaborado pelo autor (2026).

Para síntese de teste, utilizaram-se as cinco frases em português brasileiro já empregadas nos demais experimentos do projeto.[10] A avaliação recaiu sobre o *checkpoint* final (época 150) porque o salvamento de *checkpoints* foi realizado por intervalos de tempo, e não pela métrica de perda de validação; assim, o *checkpoint* correspondente à menor perda (época 136) não foi retido. Dada a pequena diferença entre as perdas (0,590 e 0,644) e o caráter de evidência negativa do experimento, essa escolha não compromete a conclusão. A avaliação informal por escuta das cinco sínteses do *checkpoint* final foi negativa: as saídas não produziram fala inteligível, sendo inutilizáveis para avaliação perceptiva de sotaque.

[10] Amostras de áudio disponíveis em `https://thomazdiniz.github.io/fala_pb/amostras/`.

Interpreta-se que o volume de dados é insuficiente para treinamento a partir de pesos aleatórios em TTS multilíngue, regime em que a literatura costuma exigir corpora da ordem de milhares de horas, a exemplo do pré-treino multilíngue do XTTS (Tabela 3.1) (Casanova et al., 2024) e bases como o TAGARELA (Oliveira et al., 2026). Essa hipótese reforça-se pelo contraste com os ajustes finos do F5-TTS sobre pesos pré-treinados (Seção 5.3.2), nos quais a síntese permaneceu utilizável, ainda que sem incorporação clara do sotaque paraibano.

Resultado e implicação. O treinamento do modelo F5-TTS mostrou que a redução da perda de validação não implica síntese audível quando o modelo é treinado a partir de pesos aleatórios com aproximadamente 111 horas de dados. Por isso, a linha de ajuste fino sobre F5-TTS pré-treinado permanece mais adequada ao escopo da dissertação; o treinamento a partir de pesos aleatórios foi documentado como evidência negativa e não seguiu para a etapa de avaliação.

5.3.4 Comparação entre Estratégias de Adaptação

As três estratégias previstas na Questão de Pesquisa 2 (Seção de questões de pesquisa no Capítulo 4) foram investigadas com custos distintos de tempo de GPU, de preparação de dados e de promessa para generalizar o sotaque paraibano a textos novos. A Tabela 5.11 resume, em termos de tempo e de desfecho qualitativo preliminar, o que foi observado nesta dissertação. Os detalhes por modelo estão nas Seções 5.3.1, 5.3.2 e 5.3.3.

Tabela 5.11: Comparação das estratégias de adaptação quanto ao tempo de computação e ao desfecho qualitativo preliminar.

Estratégia	**Tempo de GPU**	**Escopo experimental**	**Desfecho preliminar sobre sotaque / inteligibilidade**
Clonagem de fala	Desprezível	Exploratória; referências de 10–60 s	Bom timbre; sotaque não capturado em nenhuma condição
Ajuste fino	Entre 4 horas e 4 dias dependendo da arquitetura	Cinco modelos treinados sobre *FALA_PB*	Melhor equilíbrio: síntese utilizável; sotaque perceptível em parte dos casos (ex.: Orpheus TTS, Fish Speech)
Treinamento a partir de pesos aleatórios (F5-TTS)	10 dias, 150 épocas	Um único modelo treinado sobre *FALA_PB*; pesos aleatórios	Perda de validação em queda, mas síntese não inteligível; excluído das validações seguintes

Tempos de ajuste fino: XTTS 4 h 27 min; F5-TTS 5 h 13 min; Orpheus TTS 21 h 22 min; Fish Speech 44 h 43 min; Qwen3-TTS 4 dias 3 h 28 min. Clonagem sem atualização de pesos. Treinamento a partir de pesos aleatórios (F5-TTS): 244 h, aproximadamente 10 dias em 150 épocas (Seção 5.3.3).

Fonte: elaborada pelo autor (2026).

Do ponto de vista do custo computacional e da capacidade limitada da infraestrutura pessoal, a clonagem exige pouquíssimo tempo para treinamento, mas não capturou o sotaque em nenhuma condição testada, reproduzindo apenas o timbre do locutor, e foi tratada como etapa exploratória, não como linha principal de adaptação. O treinamento a partir de pesos iniciais aleatórios concentrou o maior investimento computacional em um único experimento e não produziu fala utilizável, funcionando como evidência negativa sobre a viabilidade dessa rota com aproximadamente 111 horas de corpus regional.

O ajuste fino reuniu o melhor compromisso para o escopo desta dissertação: tempo para treinamento ainda elevado quando se somam várias arquiteturas, porém ordens de magnitude

abaixo do tempo requerido para treinamento do modelo F5-TTS (pesos iniciais aleatórios), com síntese inteligível na maioria dos modelos testados e indícios de sotaque paraibano em pelo menos parte dos *checkpoints*. Por isso, as etapas de avaliação automática e perceptiva (Seções 5.4 e 5.5) priorizaram *checkpoints* obtidos por ajuste fino, e não o modelo treinado a partir de pesos aleatórios.

5.3.5 Condições Experimentais e Proporção de Dados Regionais

A comparação entre os experimentos exige que se explicitem as condições sob as quais cada um foi executado, uma vez que elas não foram uniformes. A Tabela 5.12 reúne, para cada experimento, o *checkpoint* de partida, o número de épocas de treinamento executadas, o tempo de computação e o desfecho observado. Todos os experimentos utilizaram o mesmo corpus, o *FALA_PB*, com 111 h 45 min de fala paraibana em 34.866 amostras (Tabela 5.4).

Tabela 5.12: Condições experimentais e desfecho por experimento.

Modelo	***Checkpoint* de partida**	**Épocas**	**Tempo de GPU**	**Desfecho**
XTTS	coqui/XTTS-v2 (multilíngue, PT-BR nativo)	3	4 h 27 min	Funcionou; seguiu para avaliação
F5-TTS	firstpixel/F5-TTS-pt-br (comunitário)	3	5 h 13 min	Não funcionou; síntese sem alteração audível
Orpheus TTS	ArtooDtoo/Orpheus_PTBR_FT (comunitário)	1	21 h 22 min	Funcionou; seguiu para avaliação
Fish Speech	*Checkpoint* oficial multilíngue	1	44 h 43 min	Funcionou; seguiu para avaliação
Qwen3-TTS	Qwen3-TTS-12Hz-1.7B-Base	1	4 dias 3 h 28 min	Não funcionou; excluído da avaliação
F5-TTS (pesos aleatórios)	Nenhum	150	≈244 h	Não funcionou; síntese não inteligível
Clonagem *zero-shot*	Modelos pré-treinados, pesos congelados	—	Desprezível	Não funcionou para sotaque; timbre preservado

Todos os experimentos utilizaram o corpus *FALA_PB* (111 h 45 min; Tabela 5.4). O critério de parada, definido na Seção 4.5.2, previa no máximo três épocas completas, com encerramento após a primeira época caso o treino ultrapassasse cerca de 24 horas.

Fonte: elaborada pelo autor (2026).

Três observações decorrem dessas condições. A primeira é que as condições experimentais não são uniformes e, portanto, a comparação entre modelos não constitui um experimento controlado. O critério de parada adotado, baseado em tempo de execução e não em convergência, fez com que os modelos XTTS e F5-TTS recebessem três épocas de treinamento enquanto os demais receberam apenas uma, o que confunde o efeito da arquitetura com o do orçamento de treinamento efetivamente aplicado a cada modelo. Os resultados devem ser lidos, assim, como o desempenho de cada arquitetura sob a restrição real de infraestrutura da pesquisa.

A segunda observação diz respeito aos dois experimentos de ajuste fino que não produziram efeito audível. No caso do modelo Qwen3-TTS, destaca-se a taxa de aprendizado empregada, 2×10^{-6}: o valor não foi definido nesta pesquisa, mas herdado da configuração de referência disponibilizada pelos autores do modelo (Hu et al., 2026), e é cerca de cinco vezes menor que a dos modelos XTTS e Orpheus TTS e cinquenta vezes menor que a do Fish Speech; ainda que a perda de treino tenha caído ao longo de 20.976 passos, é plausível que o deslocamento dos pesos tenha sido pequeno demais para alterar perceptivelmente a pronúncia, hipótese compatível com o fato de o modelo ter sido, ao mesmo tempo, o mais caro em tempo de GPU e o menos responsivo. No caso do modelo F5-TTS, o ajuste partiu de um *checkpoint* de domínio público já adaptado ao português, configurando um segundo ajuste em cadeia sobre pesos previamente deslocados; os arquivos de pesos mudaram, mas a síntese permaneceu praticamente idêntica, o que sugere que a atualização se concentrou em componentes sem efeito sobre os traços regionais.

A terceira observação refere-se à escala do corpus regional acrescentado, quando comparada ao volume de dados sobre o qual os modelos foram originalmente pré-treinados. Essa comparação só é possível para o XTTS, único modelo cujo corpus de pré-treinamento é reportado de forma verificável (Casanova et al., 2024): as 111,8 horas do corpus *FALA_PB* correspondem a aproximadamente 4,7% das 2.386,8 horas de português e a menos de 0,5% das 27.281,6 horas totais utilizadas no pré-treinamento (Tabela 3.1). O sinal regional introduzido pelo ajuste fino é, portanto, minoritário mesmo dentro do idioma alvo, o que oferece uma explicação plausível para o sotaque misto observado nas amostras de acompanhamento (Seção 5.3.2) e para o fato de a incorporação, ainda que perceptível aos ouvintes, não substituir integralmente o padrão de pronúncia herdado do *checkpoint* de origem. Para os demais

modelos, a composição do corpus de pré-treinamento não é documentada de forma verificável, o que impede o mesmo cálculo e restringe a comparação direta entre as arquiteturas.

5.3.6 Discussão sobre a Questão de Pesquisa 2

A Questão de Pesquisa 2 pergunta quais estratégias de adaptação são viáveis para incorporar o sotaque paraibano, considerando os recursos disponíveis. Com os recursos disponíveis nesta pesquisa (uma única estação com GPU, tempo de computação limitado e corpus regional na ordem de uma centena de horas de fala), a resposta apoiada nos experimentos é que o ajuste fino de modelos já pré-treinados é, de longe, a alternativa mais adequada entre as três investigadas.

A clonagem permanece útil como baseline de baixo custo e para imitar timbre com poucos segundos de referência, mas não capturou o sotaque em nenhuma das condições testadas. O treinamento a partir de pesos iniciais aleatórios, embora conceitualmente atraente para "reaprender" o português regional a partir do *FALA_PB*, demandou cerca de dez dias de GPU em uma única arquitetura e não gerou síntese audível; este fato reforça que escalas de pré-treino da ordem de milhares de horas (Tabela 3.1; (Oliveira et al., 2026)) não foram reproduzíveis neste trabalho. Entre os ajustes finos, os desfechos variaram por modelo (XTTS e F5-TTS com sotaque ainda tímido; Orpheus TTS e Fish Speech com percepção mais favorável na escuta informal; Qwen3-TTS com perda, *loss*, em queda, porém sem ganho audível de sotaque), o que confirma a hipótese de que o esforço computacional e o comportamento da arquitetura importam tanto quanto o corpus regional.

Todas essas conclusões intermediárias sobre sotaque e qualidade basearam-se em avaliação preliminar por escuta do autor desta dissertação (frases de acompanhamento e amostras pontuais após treinamento ou inferência). Isso constitui uma ameaça à validade do tipo viés de expectativa e falta de generalização para outros ouvintes: julgamentos não foram obtidos de forma anônima, balanceada nem por meio de avaliação perceptiva formal com participantes externos. O desenho foi deliberado para reduzir custo e tempo durante o ciclo intensivo de treinos, mas não substitui avaliação perceptiva formal.

As próximas subseções deste capítulo endereçam essa limitação a um custo muito menor do que o de novos treinamentos de grande porte: métricas automáticas de qualidade e similaridade textual (Questão de Pesquisa 3) e, na avaliação com participantes, as compara-

ções perceptivas de sotaque e de clareza/naturalidade previstas na metodologia (Questão de Pesquisa 4). Esses instrumentos permitem comparar *checkpoints* de forma mais sistemática e reprodutível, sem repetir dezenas de dias de GPU. O treinamento a partir de pesos aleatórios do modelo F5-TTS e o ajuste fino do Qwen3-TTS, sem ganho perceptivo, ilustram, adicionalmente, que métricas de otimização (perda de treino ou de validação) não devem ser confundidas com adequação para a Questão de Pesquisa 2 sem confirmação perceptiva com voluntários.

Em síntese, para incorporar sotaque paraibano sob restrição real de infraestrutura e de escala de dados, esta dissertação indica o ajuste fino como estratégia central; a clonagem como complemento exploratório de baixo custo; e o treinamento a partir de pesos iniciais aleatórios, no estado atual do corpus, como rota não viável no prazo e nos recursos da pesquisa. A validação definitiva da Questão de Pesquisa 2 repousa nas avaliações automáticas e subjetivas apresentadas nas Seções 5.4 e 5.5, e não na escuta exclusiva dos autores.

Dentre os cinco modelos ajustados, apenas o XTTS, o Fish Speech e o Orpheus TTS seguiram para as avaliações automática e perceptiva, por terem produzido sínteses utilizáveis e com indícios observáveis de adaptação regional. O F5-TTS não apresentou alteração audível relevante após o ajuste fino e o Qwen3-TTS, apesar da redução da perda de treino, não evidenciou aproximação perceptível ao sotaque paraibano; ambos foram, por isso, mantidos fora das comparações das próximas subseções. O quadro seguinte resume a resposta dada à Questão de Pesquisa 2.

Questão de Pesquisa 2: Quais estratégias de adaptação são viáveis para a incorporação do sotaque paraibano em modelos de síntese de fala, considerando os recursos computacionais e de dados disponíveis?

Resposta: O ajuste fino de modelos pré-treinados é a estratégia mais adequada sob as restrições reais de infraestrutura e de escala de dados desta pesquisa. A clonagem de fala reproduziu bem o timbre, mas não capturou o sotaque em nenhuma condição testada, permanecendo como recurso exploratório de baixo custo; o treinamento a partir de pesos aleatórios, com aproximadamente 111 horas de corpus, não produziu fala inteligível após cerca de dez dias de GPU, constituindo evidência negativa sobre essa rota. Entre os cinco modelos ajustados, XTTS, Fish Speech e Orpheus TTS apresentaram indícios perceptíveis de adaptação regional e seguiram para as avaliações automática e perceptiva (Seções 5.4 e 5.5), que confirmaram a viabilidade da estratégia.

5.4 Resultados das Avaliações Automáticas

A Questão de Pesquisa 3 investiga se os ajustes realizados nos modelos de síntese de fala impactam negativamente a qualidade e a inteligibilidade das falas sintetizadas. Para respondê-la de forma reprodutível e independente de ouvintes humanos, esta seção apresenta os resultados da avaliação automática conduzida sobre os áudios sintetizados pelos modelos adaptados por ajuste fino (Seção 5.3.2), comparando cada checkpoint ajustado (com sotaque paraibano) ao respectivo checkpoint original (modelo base, antes da adaptação). O protocolo segue o que foi definido na Seção 4.6.

A comparação abrange os três modelos adaptados nesta pesquisa: Fish Speech, Orpheus TTS e XTTS, totalizando 600 áudios sintetizados[11] conforme o protocolo definido na Seção 4.6 (100 sentenças de teste em duas condições, original e ajustada, para cada um dos três modelos).

[11] Amostras de áudio disponíveis em `https://thomazdiniz.github.io/fala_pb/amostras/`.

5.4.1 Métricas e Procedimento de Agregação

Duas famílias de métricas objetivas, ambas fundamentadas no Capítulo 2 e na Seção 4.6, foram empregadas:

- Similaridade de Levenshtein entre o texto enviado ao sintetizador (*ground truth*) e a transcrição obtida por reconhecimento automático de fala (Whisper, modelo *small*, idioma português). A similaridade é normalizada em nível de caractere, no intervalo $[0, 1]$, em que $1{,}0$ indica transcrição idêntica. Esta é a métrica central de inteligibilidade: valores baixos indicam que o conteúdo lexical da fala não foi corretamente recuperado pelo reconhecedor.

- Preditores automáticos de MOS, não intrusivos (sem áudio de referência nem ouvintes), nas implementações *NISQA* (Mittag et al., 2021) e *DNSMOS* (Reddy et al., 2022). Do *NISQA* aproveitam-se a qualidade geral (*overall*) e as dimensões de ruído percebido (*noisiness*), descontinuidade (*discontinuity*), coloração (*coloration*) e intensidade sonora (*loudness*). Do *DNSMOS* aproveitam-se as estimativas de qualidade do sinal de voz (*sig*), do fundo/ruído de fundo (*bak*) e da qualidade global (*ovrl*). Nesta rodada, a coluna de MOS principal foi atribuída ao *NISQA overall*.

Conforme o protocolo definido na Seção 4.6, cada sentença gera uma observação por condição, de modo que o teste pareado incide sobre 100 pares original–ajustado por modelo.

A significância das diferenças entre condições foi avaliada pelo teste de postos sinalizados de Wilcoxon (Wilcoxon, 1945), pareado por índice de sentença, teste não paramétrico adequado a escores limitados e de distribuição assimétrica como os dos preditores de MOS (Seção 2.7.5). Em todas as tabelas a seguir, define-se $\Delta = \overline{\text{ajustado}} - \overline{\text{original}}$, de modo que $\Delta > 0$ indica vantagem do modelo ajustado; marca-se com asterisco (*) o resultado com $p < 0{,}05$. Para cada métrica são apresentados três tipos de visualização, todos disponíveis no Apêndice C para o conjunto completo de métricas: gráficos de barras (médias por modelo e condição), diagramas de caixa (*boxplots*, distribuição e dispersão por modelo e condição) e histogramas (frequência dos valores por modelo e condição).

5.4.2 Inteligibilidade: Similaridade de Levenshtein

A similaridade de Levenshtein é a métrica mais diretamente associada à pergunta "o ouvinte entende o que foi dito?". A Tabela 5.13 sintetiza os resultados.

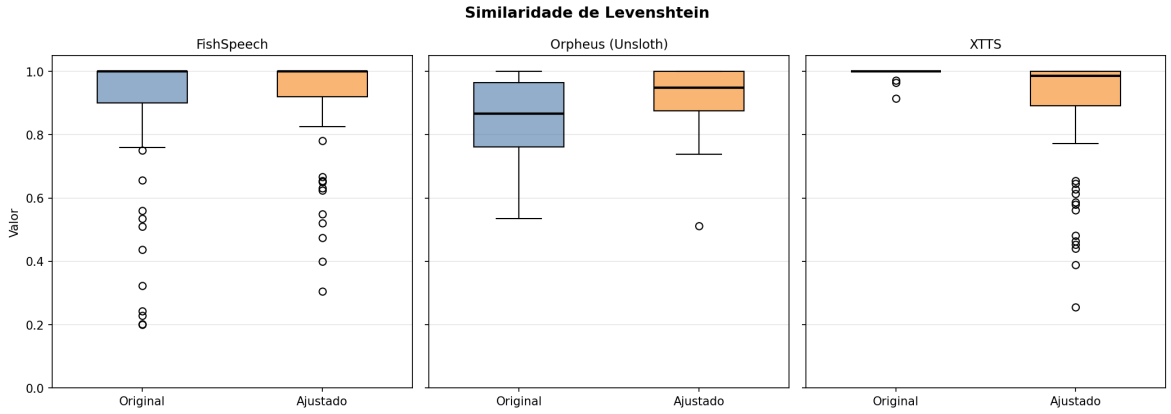
Tabela 5.13: Similaridade de Levenshtein (inteligibilidade) por modelo: original vs. ajustado.

Modelo	**Original**	**Ajustado**	Δ	p	**Leitura**
Fish Speech	0,906	0,927	$+0{,}021$	0,453	sem diferença significativa (leve melhora)
Orpheus TTS	0,855	0,921	$+0{,}066^{*}$	0,001	ajustado melhor
XTTS	0,999	0,904	$-0{,}095^{*}$	$< 0{,}001$	queda pequena, nível ainda alto

Fonte: elaborada pelo autor (2026).

A Figura 5.11 apresenta a distribuição dessas similaridades por modelo e condição. Nos três modelos, as caixas das condições original e ajustada ocupam a mesma faixa superior da escala, próximas de $1{,}0$: nos *boxplots* dos modelos Fish Speech e Orpheus TTS, os modelos ajustados se igualam ou superam os modelos originais. No modelo XTTS, embora a caixa do ajustado desça um pouco em relação ao original (quase uma linha em $1{,}0$), ela permanece concentrada no topo, com apenas alguns pontos de cauda inferior. Embora o XTTS apresente redução estatisticamente significativa na similaridade de Levenshtein (de $0{,}999$ para $0{,}904$; Tabela 5.13), o valor permanece em patamar alto e a distribuição segue concentrada próxima de $1{,}0$, o que sugere preservação prática da inteligibilidade; por outro lado, nos modelos Fish Speech e Orpheus TTS não há perda. No conjunto, a adaptação do sotaque não produziu perda relevante de inteligibilidade. O conjunto completo de gráficos (barras, caixas e histogramas) está no Apêndice C.

Figura 5.11: Similaridade de Levenshtein: distribuição por modelo e condição (azul: original; laranja: ajustado).



Fonte: elaborado pelo autor (2026).

5.4.3 Qualidade Perceptiva Global: *DNSMOS overall*

O *DNSMOS overall* é um preditor de qualidade perceptiva global robusto e metodologicamente independente do *NISQA*, o que o torna a evidência mais forte sobre (não) degradação de qualidade. A Tabela 5.14 apresenta os resultados deste preditor.

Tabela 5.14: *DNSMOS overall* (qualidade perceptiva global) por modelo: original vs. ajustado.

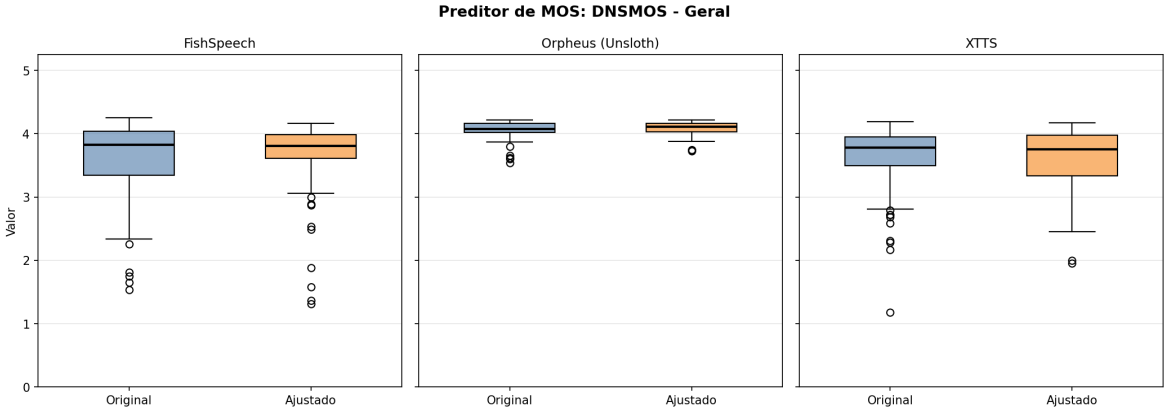
Modelo	**Original**	**Ajustado**	Δ	p	**Leitura**
Fish Speech	3,618	3,658	$+0{,}040$	0,639	sem diferença significativa
Orpheus TTS	4,056	4,072	$+0{,}016$	0,577	sem diferença significativa
XTTS	3,625	3,597	$-0{,}028$	0,680	sem diferença significativa

Fonte: elaborada pelo autor (2026).

Este é o resultado mais determinante para a Questão de Pesquisa 3. A Figura 5.12 mostra que, nos três modelos, as caixas das condições original e ajustada praticamente coincidem. Elas ocupam a mesma região da escala, têm medianas próximas e se sobrepõem bastante, o que é compatível com a ausência de diferença estatística no teste pareado. A qualidade

global percebida do áudio foi preservada após o ajuste fino.

Figura 5.12: *DNSMOS overall*: distribuição por modelo e condição (azul: original; laranja: ajustado).



Fonte: elaborado pelo autor (2026).

5.4.4 MOS Principal: *NISQA overall*

O preditor escolhido como MOS principal nesta rodada foi o *NISQA overall*. A Tabela 5.15 apresenta os resultados; a leitura aqui exige cuidado adicional, discutido em seguida.

Tabela 5.15: *NISQA overall* (MOS principal) por modelo: original vs. ajustado.

Modelo	**Original**	**Ajustado**	Δ	p	**Leitura**
Fish Speech	3,645	3,563	$-0{,}082^{*}$	0,031	queda marginal ($\sim$2%)
Orpheus TTS	4,846	4,252	$-0{,}594^{*}$	$< 0{,}001$	queda
XTTS	3,211	3,432	$+0{,}221^{*}$	$< 0{,}001$	ajustado melhor

Fonte: elaborada pelo autor (2026).

A leitura por modelo é a seguinte. No modelo XTTS, o ajuste fino melhora significativamente o MOS, deslocando a distribuição para cima (Figura C.5, Apêndice C). No modelo Fish Speech, a queda é de apenas 0,08 ponto ($\sim$2% em uma escala que vai até 5). O Orpheus TTS é o único modelo que apresentou queda expressiva no *NISQA*, partindo de um patamar próximo do teto da escala ($\sim$4,85) no *checkpoint* original.

5.4.5 Dimensões de Qualidade: Ruído ou Mudança de Sotaque?

A Tabela 5.16 reúne as dimensões mais finas do *NISQA* e do *DNSMOS*, que permitem interpretar a natureza das variações observadas.

Tabela 5.16: Dimensões de qualidade (*NISQA* e *DNSMOS*) por modelo: $\Delta = \overline{\text{ajustado}} - \overline{\text{original}}$ (*: $p < 0{,}05$).

Dimensão	**Fish Speech**	**Orpheus TTS**	**XTTS**
NISQA noisiness (ruído)	$-0{,}201^*$	$-0{,}795^*$	$-0{,}456^*$
NISQA discontinuity (descontinuidade)	$-0{,}072$	$-0{,}231^*$	$+0{,}576^*$
NISQA coloration (coloração)	$-0{,}090$	$-0{,}192^*$	$+0{,}338^*$
NISQA loudness (intensidade)	$-0{,}090^*$	$-0{,}315^*$	$+0{,}357^*$
DNSMOS sig (sinal de voz)	$+0{,}007$	$-0{,}160^*$	$+0{,}055^*$
DNSMOS bak (fundo)	$-0{,}021$	$+0{,}078^*$	$+0{,}219^*$

Fonte: elaborada pelo autor (2026).

A única dimensão que cai consistentemente nos três modelos é a *NISQA noisiness* (Fish Speech $-0{,}20$; Orpheus TTS $-0{,}80$; XTTS $-0{,}46$). As demais métricas de ruído, no entanto, não acompanham essa queda:

- O *DNSMOS bak* (qualidade do fundo, isto é, presença de ruído de fundo) não piora: fica estável nas sínteses do modelo Fish Speech e melhora significativamente nas sínteses dos modelos Orpheus TTS ($+0{,}08$) e XTTS ($+0{,}22$). Se houvesse ruído aditivo real introduzido pela adaptação, o *bak* cairia.
- O *DNSMOS overall* permanece estável nos três modelos (Seção 5.4.3).

A queda de *noisiness* não é acompanhada de piora na qualidade global: o *DNSMOS overall* e o *bak* permanecem estáveis ou melhoram, o que torna improvável que ela corresponda a ruído introduzido pela adaptação. Uma explicação mais plausível está nas características acústicas herdadas do corpus de adaptação, tanto nas condições de gravação (microfone e ambiente de captação das amostras paraibanas) quanto nos traços espectrais e prosódicos do próprio sotaque. Essas características se refletem nas saídas do modelo adaptado e o *NISQA* tende a penalizá-las nessa dimensão. No modelo XTTS, a adaptação ainda veio acompanhada de melhora em descontinuidade, coloração e intensidade. No modelo Orpheus TTS,

o fundo ficou mais limpo (*bak* $+0{,}08$) e a inteligibilidade aumentou (Levenshtein $+0{,}066$), um perfil mais compatível com mudança de timbre do que com perda de qualidade.

5.4.6 Discussão sobre a Questão de Pesquisa 3

Os indicadores objetivos não apontam degradação global consistente após o ajuste fino para incorporação do sotaque paraibano. Algumas métricas específicas, como o *NISQA overall* no Orpheus TTS e a similaridade de Levenshtein no XTTS, apresentam variações que exigem interpretação cautelosa. Em resumo:

- Inteligibilidade sem perda relevante: a similaridade de Levenshtein sobe em Fish Speech e Orpheus TTS (significativa neste último); no XTTS há queda estatisticamente significativa, mas o valor permanece em patamar alto ($0{,}90$), compatível com preservação prática da inteligibilidade.

- Qualidade global intacta: o *DNSMOS overall* não apresenta diferença significativa em nenhum dos três modelos.

- As variações observadas por modelo, ainda que algumas estatisticamente significativas (como a similaridade de Levenshtein no XTTS e o *NISQA overall* no Orpheus TTS), são de pequena magnitude prática e não se confirmam como perda de qualidade ou de inteligibilidade na percepção humana (Seção 5.5).

- A queda de *noisiness* não corresponde a piora da qualidade global (*DNSMOS overall* e *bak* estáveis ou melhores) e pode refletir características acústicas herdadas do corpus de adaptação (condições de gravação e traços regionais), e não ruído introduzido pela adaptação.

Cabe a ressalva metodológica de que os preditores automáticos de MOS são estimativas relativas, úteis para comparar checkpoints sob o mesmo *pipeline*, e não substituem o MOS subjetivo da norma ITU-T P.800 (ITU-T, 1996). A confirmação definitiva da Questão de Pesquisa 3, e em especial da hipótese de que as variações detectadas correspondem ao sotaque e não a perda de qualidade, vem da avaliação perceptiva com ouvintes, apresentada na Seção 5.5, cujos resultados são plenamente convergentes com os indícios objetivos

aqui reportados. O quadro seguinte sintetiza os achados dos experimentos reportados nesta subseção atinentes à Questão de Pesquisa 3.

Limitações das Métricas Objetivas

Cabe explicitar que nenhuma das métricas objetivas empregadas nesta seção mede o sotaque. A similaridade de Levenshtein afere a preservação do conteúdo lexical, isto é, verifica se as palavras enviadas ao sintetizador podem ser recuperadas a partir do áudio gerado; o *NISQA* e o *DNSMOS* estimam qualidade perceptual e ruído, tendo sido treinados para prever julgamentos de qualidade sobre fala majoritariamente neutra. Nenhuma delas dispõe de qualquer representação de variedade regional. Em consequência, a avaliação automática responde à Questão de Pesquisa 3, relativa a eventual degradação de qualidade e de inteligibilidade, mas é, por construção, incapaz de responder à Questão de Pesquisa 4, relativa à incorporação do sotaque, cuja verificação depende integralmente da avaliação subjetiva apresentada na Seção 5.5. Quando, ao longo desta seção, variações em dimensões como o *noisiness* são associadas a traços regionais, trata-se de hipótese interpretativa formulada *a posteriori*, e não de medição direta do sotaque.

Possíveis Causas das Diferenças entre os Modelos

As variações observadas entre os três modelos possivelmente não são de mesma natureza e admitem explicações distintas, que se apresentam aqui como hipóteses. A primeira diz respeito ao ponto de partida de cada *checkpoint*: o modelo XTTS original já operava praticamente no teto da escala de inteligibilidade ($0{,}999$), de modo que qualquer deslocamento produzido pelo ajuste só poderia ocorrer para baixo, ao passo que o modelo Orpheus TTS partia do patamar mais baixo ($0{,}855$) e dispunha, portanto, de margem para melhora. Esse efeito de teto explica boa parte da assimetria observada nos sinais dos Δ, sem que seja necessário supor perda de qualidade no XTTS.

A segunda hipótese relaciona-se às condições de cada ajuste fino, descritas na Seção 5.3.2. O *checkpoint* base do modelo Orpheus TTS apresentava artefatos de síntese, e o ajuste sobre um corpus regional padronizado e filtrado por qualidade acústica pode ter estabilizado a pronúncia, o que é coerente com o único ganho estatisticamente significativo de inteligibilidade observado. O XTTS, por sua vez, foi o único modelo ajustado por três épo-

cas, contra uma época nos demais, o que implica maior deslocamento dos pesos em relação ao *checkpoint* de origem e, plausivelmente, maior afastamento do padrão de fala neutra sobre o qual as métricas automáticas foram calibradas.

A terceira hipótese, e possivelmente a mais relevante, diz respeito ao instrumento de medida da inteligibilidade. A similaridade de Levenshtein é calculada sobre transcrições produzidas pelo modelo Whisper (Radford et al., 2023), um reconhecedor de fala treinado majoritariamente sobre variedades padronizadas do português. É razoável supor que sua taxa de erro seja mais alta justamente sobre a fala em que os traços regionais são mais salientes, de modo que a queda observada no modelo XTTS pode refletir o próprio sucesso da adaptação, e não uma perda real de inteligibilidade. Essa leitura é sustentada pela avaliação perceptiva: o modelo XTTS foi simultaneamente o modelo mais bem avaliado quanto à presença do sotaque e o modelo cuja clareza os ouvintes consideraram preservada, resultado incompatível com degradação efetiva do conteúdo. Essa ameaça à validade é retomada na Seção 5.6.3.

Questão de Pesquisa 3: Os ajustes realizados nos modelos de síntese de fala impactam negativamente a qualidade e a inteligibilidade das falas sintetizadas?

Resposta: Não de forma relevante. A avaliação automática sobre 600 áudios sintetizados, comparando cada modelo ajustado ao respectivo modelo original, indica que: (i) a inteligibilidade, medida pela similaridade de Levenshtein entre o texto de entrada e a transcrição automática, é preservada ou aprimorada nos três modelos (com melhora significativa no Orpheus TTS e nível ainda alto no XTTS); e (ii) a qualidade perceptiva global, medida pelo *DNSMOS overall*, não apresenta diferença estatisticamente significativa em nenhum dos três modelos. As quedas observadas em dimensões mais finas dos preditores, sobretudo o *NISQA noisiness*, podem estar associadas a características acústicas herdadas do corpus de adaptação (condições de gravação e traços espectrais e prosódicos do sotaque), e não necessariamente a degradação ou ruído, hipótese reforçada pela estabilidade ou melhora do ruído de fundo (*DNSMOS bak*). Esses indícios objetivos são confirmados pela avaliação perceptiva com ouvintes (Questão de Pesquisa 4).

5.5 Resultados das Avaliações Subjetivas

A Questão de Pesquisa 4 investiga se os ajustes realizados nos modelos de síntese de fala conseguem produzir percepção consistente do sotaque paraibano nas falas sintetizadas, bem como se essa adaptação preserva a clareza e a naturalidade percebidas. Para respondê-la, foi conduzida uma avaliação perceptiva com ouvintes humanos, complementar à avaliação automática da Seção 5.4, seguindo o protocolo definido na Seção 4.7.

5.5.1 Estrutura da Avaliação e Perfil dos Participantes

A avaliação foi aplicada por meio de formulário *on-line*, no qual cada participante ouvia pares de áudios e, para cada par, respondia a duas perguntas: (i) qual áudio representa melhor o sotaque paraibano; e (ii) considerando clareza e naturalidade, qual áudio é melhor. As opções de resposta incluíam, além de "Áudio A" e "Áudio B", as alternativas "Ambos aparentam ter sotaque paraibano"/"Ambos soam igualmente bem" e "Não consigo distinguir". A ordem de apresentação A/B foi aleatorizada por questão, de modo que a posição não revela qual áudio corresponde a qual condição ou modelo, reduzindo viés de ordenação.

O formulário organizou-se em dois blocos:

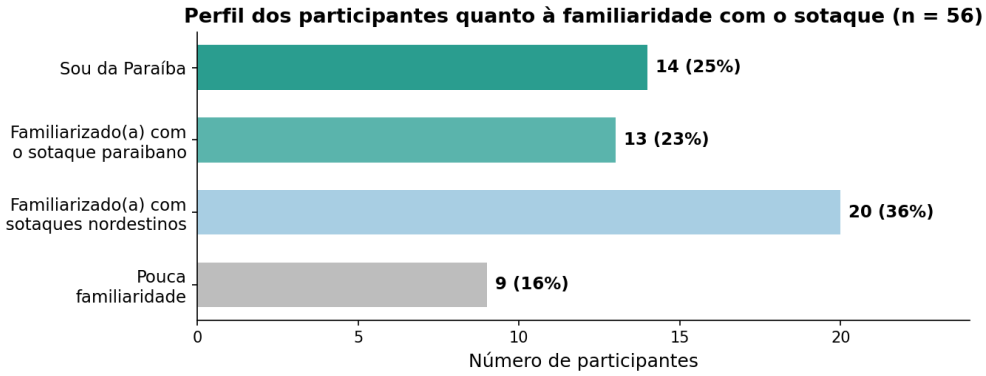
- Bloco 1 (Original $\times$ Adaptado): para cada modelo (XTTS, Fish Speech e Orpheus TTS), três amostras de áudio confrontam a versão original (base) com a versão adaptada por ajuste fino, totalizando 3 amostras $\times$ 56 respondentes $=$ 168 respostas por critério e por modelo.

- Bloco 2 (Ranqueamento entre os modelos adaptados): confrontos diretos par a par entre as versões adaptadas dos três modelos (XTTS $\times$ Fish Speech, Fish Speech $\times$ Orpheus TTS e XTTS $\times$ Orpheus TTS), com três amostras por confronto.

Participaram 56 respondentes. Quanto à familiaridade com o sotaque, 14 declararam ser da Paraíba (25,0%), 13 declararam familiaridade com o sotaque paraibano (23,2%), 20 declararam familiaridade com sotaques nordestinos em geral (35,7%) e 9 declararam pouca familiaridade (16,1%), conforme ilustra a Figura 5.13. Esse perfil permite analisar os resultados tanto no conjunto completo quanto em subgrupos mais exigentes (paraibanos e famili-

arizados): 83,9% da amostra possui alguma familiaridade com sotaques nordestinos e quase metade (48,2%) conhece especificamente o sotaque paraibano.

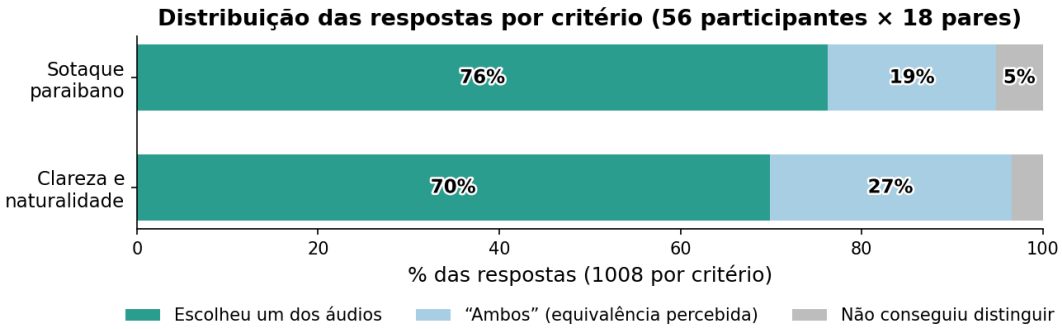
Figura 5.13: Caracterização dos participantes da avaliação perceptiva quanto à familiaridade com o sotaque paraibano ($n = 56$).



Fonte: elaborado pelo autor (2026).

Antes da análise por modelo, a Figura 5.14 caracteriza o comportamento global das respostas coletadas (1.008 por critério, correspondentes a 56 participantes e 18 pares). Na grande maioria dos itens os participantes escolheram um dos áudios (76,3% no critério de sotaque e 69,8% no de clareza e naturalidade), enquanto a opção "não consigo distinguir" foi rara (5,2% e 3,5%, respectivamente), indicando que a tarefa proposta era discriminável e que os pares de áudio apresentavam diferenças perceptíveis. As respostas de equivalência ("ambos") foram mais frequentes no critério de clareza (26,7%) do que no de sotaque (18,6%), coerente com o fato de a adaptação alterar mais a identidade sonora regional do que a inteligibilidade da fala.

Figura 5.14: Distribuição global das respostas por critério de avaliação: escolha de um dos áudios, equivalência percebida (“ambos”) e indistinguibilidade.



Fonte: elaborado pelo autor (2026).

Como o objetivo do trabalho é verificar se a fala adaptada carrega sotaque paraibano (e não apenas vencer o original em uma disputa direta), adotaram-se as seguintes convenções de leitura, favoráveis ao objetivo da adaptação: na percepção de sotaque, considera-se favorável escolher o áudio adaptado ou responder “ambos têm sotaque”; na clareza/naturalidade, considera-se que “não houve piora” quando o respondente escolhe o adaptado ou responde “ambos soam igualmente bem”. A resposta “não consigo distinguir” foi tratada como neutra e excluída dos testes binomiais. A significância foi avaliada por teste binomial exato (adaptado vs. original, descartados empates e indistinguíveis; H_0: proporção $= 0{,}5$), escolhido porque, após a exclusão dos neutros, cada resposta constitui um ensaio de Bernoulli com dois desfechos e probabilidade $0{,}5$ sob a hipótese de indistinguibilidade, de modo que a contagem de respostas favoráveis segue uma distribuição binomial (Seção 2.7.5) (Agresti, 2013). No Bloco 2, reporta-se a taxa de vitória (vitórias sobre vitórias mais derrotas, empates excluídos).

5.5.2 Percepção de Sotaque

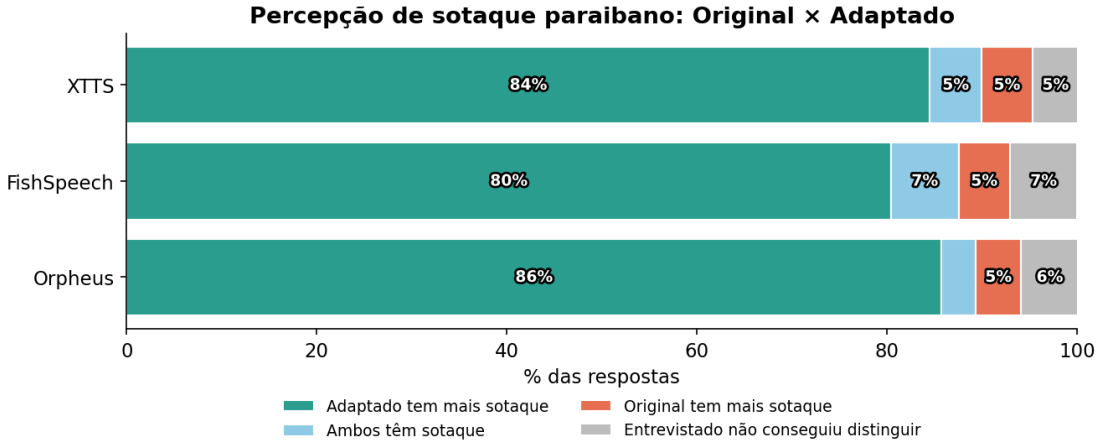
A Tabela 5.17 e a Figura 5.15 apresentam, por modelo, a distribuição das respostas sobre qual áudio carrega melhor o sotaque paraibano.

Tabela 5.17: Percepção de sotaque paraibano (Bloco 1, todos os respondentes, $n = 56$; 168 respostas por modelo).

Modelo	**Adaptado**	**Ambos**	**Original**	**Não dist.**	**Favorável**	**Binomial**
XTTS	84,5%	5,4%	5,4%	4,8%	**89,9%**	$p \approx 3{,}3 \times 10^{-32}$
Fish Speech	80,4%	7,1%	5,4%	7,1%	**87,5%**	$p \approx 2{,}7 \times 10^{-30}$
Orpheus TTS	85,7%	3,6%	4,8%	6,0%	**89,3%**	$p \approx 1{,}1 \times 10^{-33}$

Fonte: elaborada pelo autor (2026).

Figura 5.15: Percepção de sotaque paraibano (original × adaptado) por modelo. A categoria "adaptado tem mais sotaque" (rotulada na legenda) domina amplamente os três modelos.



Fonte: elaborado pelo autor (2026).

Nos três modelos, a fala adaptada foi apontada como a de sotaque paraibano em cerca de 88% a 90% das respostas favoráveis (89,9% no XTTS, 87,5% no Fish Speech e 89,3% no Orpheus TTS). Esses percentuais corroboram a hipótese de que o ajuste fino realizado adiciona o sotaque paraibano às falas sintetizadas.

5.5.3 Clareza e Naturalidade: a Adaptação Piora a Qualidade?

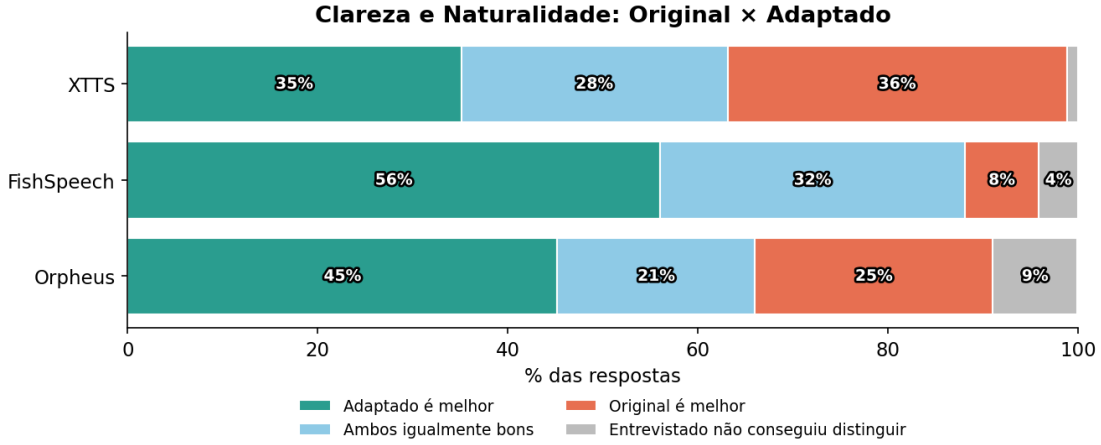
A Tabela 5.18 e a Figura 5.16 apresentam as respostas sobre clareza e naturalidade, critério perceptivo correspondente à preocupação central da Questão de Pesquisa 3 (não degradação), agora avaliado por ouvintes humanos.

Tabela 5.18: Clareza e naturalidade percebidas (Bloco 1, todos os respondentes, $n = 56$; 168 respostas por modelo).

Modelo	**Adapt.**	**Ambos**	**Orig.**	**N. dist.**	**Não piora**	**Veredito**
XTTS	35,1%	28,0%	35,7%	1,2%	**63,1%**	neutro ($p \approx 0{,}57$)
Fish Speech	56,0%	32,1%	7,7%	4,2%	**88,1%**	melhora ($p \approx 1{,}3 \times 10^{-16}$)
Orpheus TTS	45,2%	20,8%	25,0%	8,9%	**66,1%**	melhora ($p \approx 1{,}1 \times 10^{-3}$)

Fonte: elaborada pelo autor (2026).

Figura 5.16: Clareza e naturalidade percebidas (original × adaptado) por modelo. Fish Speech e Orpheus TTS apresentam preferência clara pelo adaptado; o XTTS mostra equilíbrio entre as condições.



Fonte: elaborado pelo autor (2026).

Somando as respostas favoráveis ao adaptado e as de "ambos soam igualmente bem", a clareza e a naturalidade não pioram em nenhum dos modelos: 63,1% no XTTS, 88,1% no Fish Speech e 66,1% no Orpheus TTS. Como esses percentuais ficam acima de 50%, os resultados corroboram a hipótese de que a adaptação não piora a clareza e a naturalidade percebidas.

5.5.4 Comparação Estatística entre os Modelos no Bloco 1

Os testes binomiais das seções anteriores avaliam cada modelo contra a sua própria versão original, mas não dizem se as diferenças de percentuais entre os modelos são estatisticamente significativas (por exemplo, se os 88,1% do Fish Speech em clareza são de fato maiores que

os 63,1% do XTTS, ou se a diferença é compatível com flutuação amostral). Para responder a essa pergunta, as taxas favoráveis dos três modelos foram comparadas par a par pelo teste exato de Fisher sobre tabelas de contingência 2×2 (favorável vs. não favorável, 168 respostas por modelo), com os p-valores das três comparações de cada critério ajustados pela correção de Holm (Seção 2.7.5) (Agresti, 2013; Holm, 1979). A Tabela 5.19 apresenta os resultados.

Tabela 5.19: Comparações par a par entre os modelos (Bloco 1): teste exato de Fisher com correção de Holm sobre as taxas favoráveis.

Comparação	**Critério**	**Taxas**	p **(Holm)**	**Leitura**
Fish Speech vs. XTTS	clareza	88,1% vs. 63,1%	$3{,}5\times10^{-7}$	Fish Speech maior
Fish Speech vs. Orpheus TTS	clareza	88,1% vs. 66,1%	$4{,}2\times10^{-6}$	Fish Speech maior
XTTS vs. Orpheus TTS	clareza	63,1% vs. 66,1%	0,65	sem diferença sig.
XTTS vs. Fish Speech	sotaque	89,9% vs. 87,5%	1,00	sem diferença sig.
Fish Speech vs. Orpheus TTS	sotaque	87,5% vs. 89,3%	1,00	sem diferença sig.
XTTS vs. Orpheus TTS	sotaque	89,9% vs. 89,3%	1,00	sem diferença sig.

Fonte: elaborada pelo autor (2026).

Na percepção de sotaque, as taxas favoráveis dos três modelos (89,9%, 87,5% e 89,3%) são estatisticamente indistinguíveis: os três ajustes finos adicionam o sotaque com eficácia equivalente aos olhos dos ouvintes. Na clareza e naturalidade, o quadro é diferente: a taxa de "não piora" do Fish Speech (88,1%) é significativamente maior que as do XTTS e do Orpheus TTS, enquanto XTTS (63,1%) e Orpheus TTS (66,1%) não diferem entre si. Portanto, a vantagem do Fish Speech em clareza no Bloco 1 é um efeito estatisticamente confirmado, e não flutuação amostral. Cabe a mesma ressalva de dependência intra-avaliador discutida na Seção 5.5.7: as 168 respostas por modelo agregam três respostas de cada participante, de modo que os p-valores devem ser lidos como aproximações sob suposição de independência; a convergência com a análise conservadora por participante, na qual o Fish Speech também é o único com melhora significativa de clareza, reforça a robustez da conclusão.

5.5.5 Ranqueamento entre os Modelos Adaptados

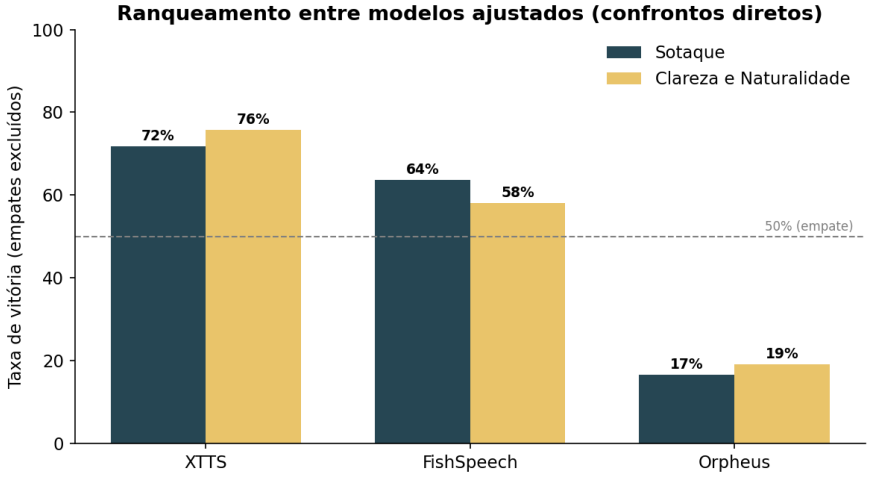
O Bloco 2 confronta diretamente as três versões adaptadas entre si. A Tabela 5.20 apresenta as taxas de vitória agregadas e a Figura 5.17 as ilustra; a Figura 5.18 detalha os confrontos par a par.

Tabela 5.20: Ranqueamento dos modelos adaptados por taxa de vitória (Bloco 2, todos os respondentes; empates excluídos).

Posição	**Modelo**	**Sotaque**	**Clareza/Naturalidade**
1º	**XTTS**	**71,8%**	**75,7%**
2º	Fish Speech	63,7%	58,1%
3º	Orpheus TTS	16,6%	19,1%

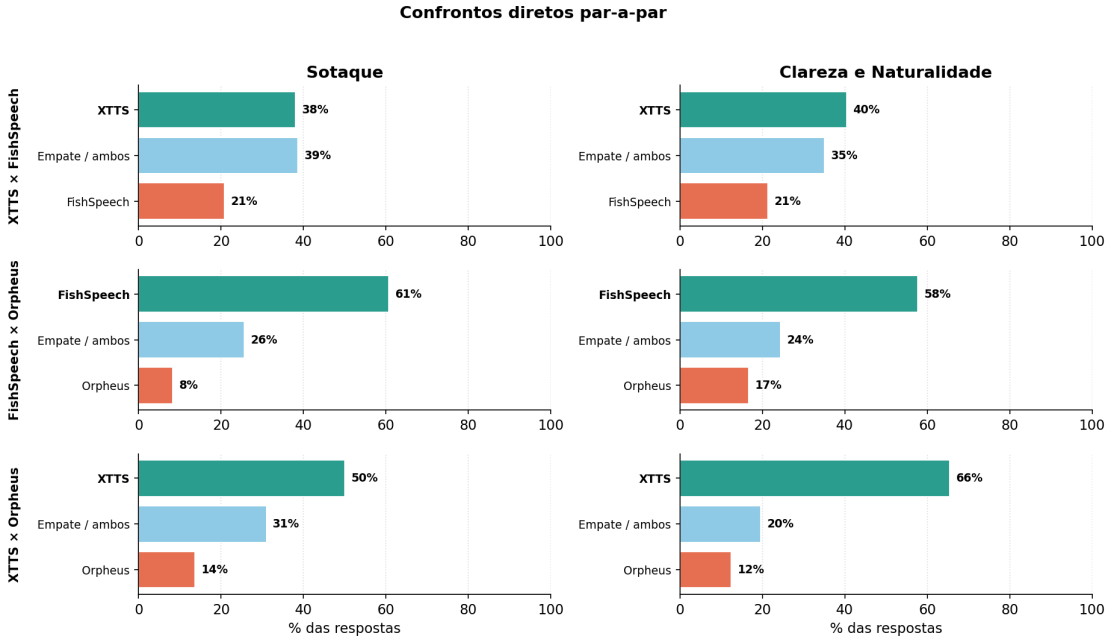
Fonte: elaborada pelo autor (2026).

Figura 5.17: Taxa de vitória de cada modelo adaptado em sotaque e em clareza/naturalidade (Bloco 2). O XTTS lidera em ambos os critérios, seguido por Fish Speech e Orpheus TTS.



Fonte: elaborado pelo autor (2026).

Figura 5.18: Confrontos diretos par a par entre os modelos adaptados, em sotaque e em clareza/naturalidade.



Fonte: elaborado pelo autor (2026).

O ranqueamento é consistente nos dois critérios: XTTS em primeiro, Fish Speech em segundo e Orpheus TTS em terceiro. Os confrontos diretos confirmam o ordenamento, com significância avaliada por teste binomial exato sobre os confrontos decididos de cada par (Seção 2.7.5): o XTTS vence o Fish Speech tanto em sotaque (38,1% × 20,8%; $p \approx 2{,}3 \times 10^{-3}$) quanto em clareza (40,5% × 21,4%; $p \approx 1{,}1 \times 10^{-3}$); o Fish Speech vence o Orpheus TTS com folga (sotaque 60,7% × 8,3%, $p \approx 5{,}6 \times 10^{-18}$; clareza 57,7% × 16,7%, $p \approx 2{,}1 \times 10^{-10}$); e o XTTS vence o Orpheus TTS em ambos os critérios (sotaque 50,0% × 13,7%, $p \approx 1{,}2 \times 10^{-9}$; clareza 65,5% × 12,5%, $p \approx 4{,}7 \times 10^{-16}$). Todas as posições do ranqueamento são, portanto, estatisticamente significativas. Entre os respondentes paraibanos, a vantagem do XTTS é ainda mais acentuada. O XTTS combina, portanto, a fala mais bem avaliada em sotaque e a mais bem avaliada em clareza e naturalidade, coerente com a melhora desse modelo também nas métricas objetivas (Seção 5.4.4).

5.5.6 Robustez por Subgrupos

Todas as análises foram recalculadas em três recortes: todos os respondentes ($n = 56$), apenas paraibanos ($n = 14$) e familiarizados (excluindo os de pouca familiaridade, $n =$

47). As conclusões são idênticas nos três recortes: o sotaque é percebido como adicionado nos três modelos (com efeito ainda mais forte entre paraibanos), a adaptação não piora a clareza/naturalidade e o ranqueamento XTTS $>$ Fish Speech $>$ Orpheus TTS se mantém. Essa estabilidade reforça a confiabilidade dos achados apesar do tamanho amostral modesto.

5.5.7 Análise Conservadora por Participante

O teste binomial por item agrega as três amostras de cada participante, de modo que as 168 respostas por modelo não são totalmente independentes (dependência intra-avaliador). Como verificação conservadora, recomputou-se a análise atribuindo a cada participante um único voto por modelo, dado pela escolha majoritária entre as três amostras. Isso reduz o tamanho efetivo para $n = 56$ participantes (empates internos excluídos).

Na percepção de sotaque, o resultado permanece fortemente significativo: a maioria dos participantes favorece o áudio adaptado em todos os modelos (XTTS, 51 contra 3; Fish Speech, 52 contra 3; Orpheus TTS, 54 contra 2; todos com $p < 0{,}001$ no teste binomial por participante). A percepção de que o ajuste fino adiciona o sotaque paraibano não é, portanto, artefato da contagem por item.

Na clareza e naturalidade, a leitura por participante é mais cautelosa: o Fish Speech mantém melhora significativa (42 contra 4; $p < 0{,}001$); o XTTS permanece neutro (22 contra 26; $p \approx 0{,}77$); e o Orpheus TTS, favorável no agregado por item, deixa de ser estatisticamente significativo neste recorte mais rigoroso (29 contra 18; $p \approx 0{,}07$). Interpreta-se que a adaptação não piora a clareza percebida, com melhora no Fish Speech e efeito neutro nos demais. Ainda assim, a vantagem do Orpheus TTS em clareza deve ser lida com cautela.

5.5.8 Discussão sobre a Questão de Pesquisa 4

A Questão de Pesquisa 4 pergunta se os ajustes produzem percepção consistente do sotaque paraibano. Os resultados da avaliação com 56 ouvintes respondem afirmativamente: nos três modelos, a versão ajustada foi percebida como a de sotaque paraibano na grande maioria das respostas, com o efeito mantido (ou intensificado) entre os respondentes paraibanos. Essa percepção não veio acompanhada de perda de clareza ou naturalidade. O Fish Speech e o Orpheus TTS melhoraram e o XTTS permaneceu estável, em convergência com a avaliação

automática (Questão de Pesquisa 3). No ranqueamento entre os modelos adaptados, o XTTS foi o mais bem avaliado nos dois critérios, seguido por Fish Speech e Orpheus TTS, e as conclusões se mantiveram nos três recortes da amostra. O quadro seguinte sintetiza a Questão de Pesquisa 4 e a resposta obtida pelos resultados reportados nesta subseção.

Questão de Pesquisa 4: Os ajustes realizados nos modelos de síntese de fala conseguem produzir percepção consistente do sotaque paraibano nas falas sintetizadas?
Resposta: Sim. A avaliação perceptiva com 56 ouvintes indica que o ajuste fino insere percepção consistente do sotaque paraibano nos três modelos avaliados, com favorabilidade de aproximadamente 88% a 90% e intensidade ainda maior entre respondentes paraibanos. Essa adaptação não degrada a clareza nem a naturalidade percebidas: melhora com os modelos Fish Speech e Orpheus TTS e é preservada no modelo XTTS, sem evidência de *trade-off* entre adicionar sotaque e manter qualidade. No ranqueamento entre os modelos adaptados, o modelo XTTS apresentou o melhor desempenho, tanto em sotaque quanto em clareza/naturalidade, seguido por Fish Speech e Orpheus TTS. Esses resultados confirmam, do ponto de vista humano, os indícios da avaliação automática (Questão de Pesquisa 3).

5.6 Análise Comparativa Geral

Esta seção integra os achados das avaliações automática (Seção 5.4) e subjetiva (Seção 5.5), relacionando-os às estratégias de adaptação discutidas na Seção 5.3.

5.6.1 Considerações Iniciais

Antes de confrontar os dois conjuntos de resultados, convém deixar explícito que eles não medem a mesma coisa e não são intercambiáveis. A avaliação objetiva (Seção 5.4) mede duas propriedades do sinal sintetizado: a preservação do conteúdo lexical, por meio da similaridade de Levenshtein calculada sobre transcrições automáticas, e a qualidade perceptual estimada, por meio dos preditores *NISQA* e *DNSMOS*. Nenhuma dessas métricas incorpora qualquer noção de variedade regional, de modo que a avaliação objetiva é capaz de indicar se a adaptação degradou a fala, mas não se ela incorporou o sotaque pretendido. A avaliação subjetiva (Seção 5.5), por sua vez, mede exatamente aquilo que as métricas automáticas não

alcançam: o julgamento de ouvintes humanos sobre a presença do sotaque paraibano e sobre a clareza e a naturalidade da fala.

As duas avaliações são, portanto, necessárias e cumprem papéis complementares. A avaliação subjetiva é o instrumento capaz de responder à Questão de Pesquisa 4, uma vez que o sotaque é, em última instância, um fenômeno reconhecido no julgamento do ouvinte; ela é, contudo, dispendiosa, limitada em tamanho amostral e sujeita a fatores como familiaridade prévia com a variedade avaliada. A avaliação objetiva é reprodutível, aplicável em escala e independente de ouvintes, o que a torna adequada ao controle de qualidade exigido pela Questão de Pesquisa 3, mas não é diretamente associada ao objeto central da pesquisa. É a conjunção das duas que sustenta a conclusão do trabalho: sem a avaliação subjetiva não haveria evidência de que o sotaque foi incorporado; sem a avaliação objetiva não haveria verificação, em escala e sem viés de ouvinte, de que essa incorporação não se deu à custa da qualidade ou da inteligibilidade.

5.6.2 Convergência entre Avaliação Automática e Subjetiva

Os dois instrumentos de avaliação (métricas automáticas e percepção humana) apontam na mesma direção, o que reforça a confiabilidade das conclusões. A Tabela 5.21 resume essa consonância por eixo de análise.

Tabela 5.21: Convergência entre as avaliações automática e subjetiva.

Eixo	**Avaliação automática**	**Avaliação subjetiva**
Inteligibilidade / clareza	Similaridade de Levenshtein preservada ou maior (alta nos três modelos)	Clareza/naturalidade não piora: melhora Fish Speech e Orpheus TTS, preserva XTTS
Qualidade global	*DNSMOS overall* sem diferença significativa nos três modelos	Sem *trade-off* perceptível entre sotaque e qualidade
Presença de sotaque	Queda de *noisiness* com fundo estável/melhor, possivelmente ligada a características herdadas do corpus de adaptação	Sotaque percebido como adicionado em $\sim$88–90% das respostas ($p \ll 0{,}001$)
Melhor modelo	XTTS melhora no MOS principal e na maioria das dimensões	XTTS em 1º lugar em sotaque e em clareza/naturalidade

Fonte: elaborada pelo autor (2026).

O ponto mais relevante dessa convergência é a interpretação da queda de *noisiness* observada na avaliação automática: a hipótese de que ela esteja associada a características acústicas herdadas do corpus de adaptação (condições de gravação e traços regionais), e não a ruído ou perda de qualidade, é reforçada (ainda que não comprovada) pela avaliação subjetiva, em que os ouvintes percebem o sotaque como fortemente presente sem reportar perda de clareza ou naturalidade.

5.6.3 Ameaças à Validade

Algumas limitações devem ser registradas para a leitura adequada dos resultados. Os preditores automáticos de MOS (*NISQA* e *DNSMOS*) são estimativas relativas, treinadas majoritariamente em fala neutra, o que torna determinadas dimensões (notadamente o *NISQA overall* e o *noisiness*) sensíveis a características de sotaque mesmo na ausência de perda real de qualidade, o que reforça a importância de interpretá-los em conjunto, e não isoladamente. Soma-se a essas limitações uma ameaça específica ao indicador de inteligibilidade: a similaridade de Levenshtein depende de transcrições produzidas pelo modelo Whisper (Radford et al., 2023), treinado majoritariamente sobre variedades padronizadas do português e sem qualquer garantia documentada de exposição à fala paraibana, cuja taxa de erro pode ten-

der, por isso, a ser mais alta justamente sobre a fala em que os traços regionais são mais salientes. Um modelo bem-sucedido em incorporar o sotaque pode, por esse motivo, ser penalizado na métrica, o que pode associar a queda observada no XTTS a um artefato do instrumento de medida, e não necessariamente a uma perda de inteligibilidade. Na avaliação subjetiva, o tamanho amostral ($n = 56$) é modesto, embora os principais efeitos sejam de grande magnitude e robustos entre subgrupos. Por fim, o corpus *FALA_PB* (Seção 5.2), apesar de adequado aos experimentos, não representa estatisticamente toda a variabilidade do sotaque paraibano, o que delimita o alcance das generalizações sobre o sotaque como um todo.

A comparabilidade entre os modelos também é limitada. Devido às limitações de *hardware* para os experimentos, as condições de treinamento não foram uniformes, conforme detalhado na Seção 5.3.5: o critério de parada, definido por tempo de execução e não por convergência, resultou em três épocas de ajuste para alguns modelos e apenas uma para outros, com número de atualizações de pesos e taxas de aprendizado igualmente distintos entre as arquiteturas. Some-se a isso o fato de o volume de fala paraibana presente no corpus construído ser pequeno em relação aos dados de pré-treinamento de cada modelo, em proporção que sequer é conhecida para a maior parte deles. A comparação entre XTTS, Fish Speech e Orpheus TTS não pode, portanto, ser considerada experimentalmente justa no sentido estrito: ela reflete o desempenho de cada arquitetura sob as condições possíveis, e diferenças observadas entre modelos podem decorrer do orçamento de treinamento tanto quanto da arquitetura em si.

Outro ponto refere-se ao estilo de fala do corpus de adaptação. O *FALA_PB* tem no CoLingPB sua principal fonte de volume (Seção 5.2), corpus originalmente constituído por entrevistas sociolinguísticas, isto é, por fala espontânea, com as hesitações, alongamentos, reformulações e variação prosódica próprios desse registro. Grande parte dos corpora empregados no pré-treinamento de sistemas TTS, ao contrário, baseia-se em leitura de sentenças, com prosódia mais controlada e regular. O ajuste fino, portanto, não apenas aproxima o modelo de uma variedade regional, mas também o desloca de um registro de fala para outro, não sendo possível separar, com os instrumentos aqui empregados, o que nas alterações observadas se deve ao sotaque e o que se deve à mudança de estilo de fala.

O material de avaliação constitui uma limitação adicional. As 100 sentenças de teste uti-

lizadas na síntese (Seção 4.6) foram definidas como conjunto fixo em português brasileiro, sem seleção orientada por critérios fonéticos. Não há garantia, assim, de que elas contenham os contextos em que os traços característicos do sotaque paraibano efetivamente se manifestam, o que pode ter reduzido a evidência disponível tanto para os ouvintes quanto para as métricas. A construção de um conjunto de sentenças foneticamente balanceado, orientado pelos traços identificados na caracterização proposta no Capítulo 6, é uma medida direta de mitigação dessa ameaça em trabalhos futuros.

5.7 Síntese do Capítulo

Este capítulo apresentou os resultados dos experimentos, organizados pelas quatro questões de pesquisa: a construção do corpus *FALA_PB* (Questão de Pesquisa 1), a comparação entre estratégias de adaptação (Questão de Pesquisa 2), a avaliação automática da qualidade e da inteligibilidade (Questão de Pesquisa 3) e a avaliação perceptiva do sotaque com ouvintes (Questão de Pesquisa 4). As avaliações automática e subjetiva foram convergentes, indicando que o ajuste fino incorporou o sotaque paraibano aos modelos investigados sem degradação relevante de qualidade ou de inteligibilidade. As implicações desses achados e as direções de continuidade são discutidas no Capítulo 6.

Capítulo 6

Conclusões e Sugestões para Trabalhos Futuros

6.1 Conclusões

Nesta dissertação foram investigados modelos de aprendizagem profunda abertos para síntese de fala, com foco na incorporação de características do sotaque paraibano ao português brasileiro. O trabalho percorreu a construção de um corpus regional, a investigação de estratégias de adaptação de modelos neurais e a avaliação automática e subjetiva das falas sintetizadas, em resposta a quatro questões de pesquisa.

Em relação à Questão de Pesquisa 1 (construção de uma base de fala voltada ao sotaque paraibano), constituiu-se o corpus *FALA_PB* por integração, filtragem e padronização de fontes complementares, com predominância do CoLingPB, e por uma coleta colaborativa complementar. O corpus mostrou-se adequado aos experimentos de adaptação realizados, ainda que não represente estatisticamente toda a variabilidade do sotaque paraibano, limitação explicitada na Seção 5.2.

Em relação à Questão de Pesquisa 2 (estratégias de adaptação), foram investigadas a clonagem de fala, o ajuste fino de modelos pré-treinados e o treinamento a partir de pesos aleatórios. Sob as restrições reais de infraestrutura e de escala de dados, o ajuste fino mostrou-se a estratégia mais adequada: a clonagem permaneceu como recurso exploratório de baixo custo e o treinamento a partir de pesos aleatórios revelou-se inviável no prazo e nos recursos disponíveis, sem ganho perceptível de sotaque.

Em relação à Questão de Pesquisa 3 (impacto na qualidade e na inteligibilidade), a avaliação automática sobre 600 áudios sintetizados indicou que o ajuste fino não degrada de forma relevante nenhuma das duas dimensões: a inteligibilidade, medida por similaridade de Levenshtein, é preservada ou aprimorada nos três modelos, e a qualidade perceptiva global, medida pelo *DNSMOS overall*, permanece estatisticamente inalterada. As variações detectadas em dimensões mais finas, sobretudo a queda de *noisiness* do *NISQA*, mostraram-se compatíveis com características acústicas herdadas do corpus de adaptação (condições de gravação e traços regionais), hipótese reforçada pela estabilidade ou melhora do ruído de fundo (*DNSMOS bak*), e não necessariamente com degradação perceptiva.

Em relação à Questão de Pesquisa 4 (percepção consistente do sotaque), a avaliação perceptiva com 56 ouvintes confirmou que o ajuste fino insere percepção consistente do sotaque paraibano nos três modelos (favorabilidade de aproximadamente 88% a 90%, com efeito altamente significativo e ainda mais forte entre respondentes paraibanos), sem prejuízo de clareza ou naturalidade. O XTTS adaptado apresentou o melhor desempenho, em sotaque e em clareza/naturalidade, seguido por Fish Speech e Orpheus TTS.

Em conjunto, os resultados automáticos e subjetivos são convergentes e sustentam a tese central do trabalho: foi possível incorporar o sotaque paraibano a modelos abertos de síntese de fala, por ajuste fino, com degradação mínima ou nula de qualidade e de inteligibilidade.

Contribuições da Pesquisa

As contribuições desta dissertação são de três naturezas, enumeradas a seguir: os artefatos abertos produzidos, os procedimentos metodológicos adotados e os resultados empíricos obtidos.

Artefatos

O primeiro conjunto de contribuições reúne três artefatos abertos, todos publicados e disponíveis para reutilização:

1. **O corpus *FALA_PB***, com cerca de 111 horas de fala paraibana segmentada em sentenças e pareada às respectivas transcrições, construído pela integração, filtragem e

padronização de fontes complementares e por coleta colaborativa própria. Encontra-se publicado na plataforma Zenodo (de Morais, 2026).

2. **Os *scripts* desenvolvidos**, que constituem o *pipeline* completo de construção do corpus, de ajuste fino dos modelos e de avaliação automática e subjetiva, permitindo a reprodução integral dos experimentos e sua reaplicação a outras variedades regionais. Encontram-se publicados em repositório aberto sob licença MIT[1].

3. **Os *checkpoints* dos modelos adaptados** ao sotaque paraibano, obtidos por ajuste fino do XTTS, do Fish Speech e do Orpheus TTS, disponíveis para uso direto em síntese de fala. Encontram-se publicados na plataforma Hugging Face[2].

Contribuições Metodológicas

O segundo conjunto refere-se aos procedimentos concebidos ou adaptados ao longo da pesquisa. São contribuições cuja aplicabilidade não se restringe ao sotaque paraibano, uma vez que descrevem como construir o dado, como medir o efeito da adaptação e como interpretar essa medida:

1. **Um procedimento de construção de corpus regional a partir de fontes heterogêneas** (Seção 4.4), que percorre a segmentação de entrevistas sociolinguísticas por marcações temporais, o corte automático, a validação por escuta, a filtragem por qualidade acústica e a padronização final do sinal, complementado por uma estratégia de coleta colaborativa de baixo custo. O procedimento é descrito em detalhe suficiente para ser reaplicado a outras variedades do português brasileiro, para as quais existam acervos sociolinguísticos comparáveis.

2. **Um instrumento de avaliação perceptiva de sotaque por comparação par a par** (Seção 4.7), submetido e aprovado pelo Comitê de Ética em Pesquisa e reproduzido na íntegra no Apêndice F. O instrumento organiza-se em dois blocos (o confronto de cada modelo com a sua própria versão adaptada e o ranqueamento das versões adaptadas entre si) e incorpora um conjunto explícito de controles de viés: apresentação da

[1] `https://github.com/ThomazDiniz/fala_pb_scripts`
[2] `https://huggingface.co/acidente/fala_pb_checkpoints`

mesma sentença nos dois áudios de cada par, anonimização dos modelos, aleatorização da posição dos áudios, inclusão de opções de equivalência e de indistinguibilidade para evitar a escolha forçada e dimensionamento do formulário para conter a fadiga do avaliador. A opção pela escala comparativa, em lugar de uma escala absoluta do tipo MOS, é justificada pela maior sensibilidade a diferenças sutis e pela redução do viés de calibração individual.

3. **Um procedimento de análise das respostas perceptivas** (Seção 4.7.9) que torna explícitas as convenções de positividade adotadas, trata as respostas inconclusivas como neutras e as exclui dos testes, e emprega o teste binomial exato para o confronto entre condições e o teste exato de Fisher com correção de Holm para as comparações entre modelos. Registre-se, em particular, a análise conservadora por participante, na qual cada respondente contribui com um único voto por modelo, adotada para neutralizar a dependência intra-avaliador introduzida pelas múltiplas amostras de um mesmo confronto, e a replicação de todas as análises em recortes por familiaridade com a variedade avaliada.

4. **Um protocolo de avaliação automática pareada por sentença** (Seção 4.6), no qual um conjunto fixo de sentenças de teste, disjunto do material de treinamento, é sintetizado nas condições original e ajustada, de modo que cada sentença contribua com um único par e o efeito da adaptação seja isolado da variabilidade do conteúdo lexical. O protocolo combina preditores de MOS (*NISQA* e *DNSMOS*) para as dimensões de qualidade perceptiva e a similaridade de Levenshtein normalizada, calculada sobre transcrições produzidas pelo modelo Whisper (Radford et al., 2023) em configuração fixa de idioma e de decodificação, como indicador de preservação do conteúdo lexical. A escolha da similaridade normalizada é justificada por operar em nível de caractere, por ser limitada e simétrica e por favorecer a leitura conjunta com os preditores de MOS; CER e WER, convenção da área, são derivados das mesmas transcrições e reportados no Apêndice D, o que preserva a comparabilidade com valores publicados por outros trabalhos.

5. **A explicitação dos limites dos instrumentos de medida empregados** (Seções 5.6.1 e 5.6.3). Nenhuma das métricas objetivas utilizadas dispõe de representação de varie-

dade regional, de modo que a avaliação automática responde pela qualidade e pela inteligibilidade, mas é incapaz, por construção, de atestar a presença do sotaque. Soma-se a isso o fato de o reconhecedor automático empregado ter sido treinado majoritariamente sobre variedades padronizadas do português, o que permite supor que sua taxa de erro seja mais alta justamente sobre a fala em que os traços regionais são mais salientes; sob essa hipótese, uma adaptação bem-sucedida pode se apresentar como queda de inteligibilidade. O registro dessa ameaça, com a recomendação de leitura conjunta das avaliações objetiva e subjetiva, é ele próprio um resultado de interesse para trabalhos que avaliem síntese de fala com sotaque.

Resultados e Conclusões

O terceiro conjunto compreende as conclusões que a pesquisa permite sustentar, obtidas a partir dos artefatos e dos procedimentos acima:

1. **Um volume moderado de fala regional é suficiente para incorporar sotaque por ajuste fino.** Cerca de 111 horas, ainda que correspondam a uma fração pequena dos dados de pré-treinamento, bastaram para produzir percepção consistente do sotaque paraibano nos três modelos avaliados, com favorabilidade entre aproximadamente 88% e 90% e efeito ainda mais acentuado entre ouvintes do próprio estado, sem prejuízo mensurável de clareza, de naturalidade ou de inteligibilidade.

2. **A clonagem *zero-shot* transfere identidade vocal, mas não sotaque.** O timbre do locutor de referência foi reproduzido de forma satisfatória, ao passo que os traços de pronúncia e de prosódia permaneceram determinados pelos pesos congelados do modelo. O resultado delimita o alcance de uma técnica frequentemente apresentada como suficiente para adaptação de voz e indica que a incorporação de sotaque exige atualização de pesos ou arquiteturas que tratem o sotaque como fator explícito.

3. **O treinamento a partir de pesos aleatórios é inviável na escala de dados e de infraestrutura considerada.** Trata-se de um resultado negativo, mas informativo, na medida em que delimita o espaço de estratégias praticáveis para grupos que disponham de recursos computacionais comparáveis aos desta pesquisa.

4. **A verificação da incorporação de sotaque depende da avaliação subjetiva, e a da qualidade, da avaliação objetiva.** Os dois instrumentos medem propriedades distintas e não são intercambiáveis; é a convergência entre eles que sustenta a conclusão central do trabalho, e não qualquer um isoladamente.

Em conjunto, os artefatos permitem que outros grupos de pesquisa verifiquem os resultados aqui relatados e retomem o trabalho a partir do ponto em que ele foi interrompido, seja ampliando o corpus, seja aplicando o mesmo procedimento a outros sotaques do português brasileiro; os procedimentos metodológicos oferecem a esses grupos um caminho já percorrido para construir o dado, medir o efeito da adaptação e interpretar essa medida com as devidas ressalvas.

6.2 Sugestões para Trabalhos Futuros

6.2.1 Aprofundamento na Caracterização do Sotaque

Um ponto importante para trabalhos futuros consiste no aprofundamento do conhecimento sobre o processo de produção e de caracterização dos sotaques, aproximando a pesquisa do instrumental já consolidado nos estudos linguísticos. Nesta dissertação, a presença do sotaque paraibano foi verificada apenas pela avaliação subjetiva dos participantes da pesquisa, sem que fossem descritos ou medidos os traços de fala que o caracterizam. Pesquisas futuras podem recorrer às métricas e às ferramentas usualmente empregadas em fonética acústica, a exemplo do *software Praat*, para verificar objetivamente esses traços em medidas como duração, taxa de elocução, formantes e contornos de frequência fundamental (Silva, 2011; Cagliari, 1992). Tal caracterização permitiria fundamentar com maior rigor as decisões metodológicas e verificar quais características o ajuste fino efetivamente transfere para a fala sintetizada.

6.2.2 Expansão da Base FALA_PB por Coleta em Redes Sociais

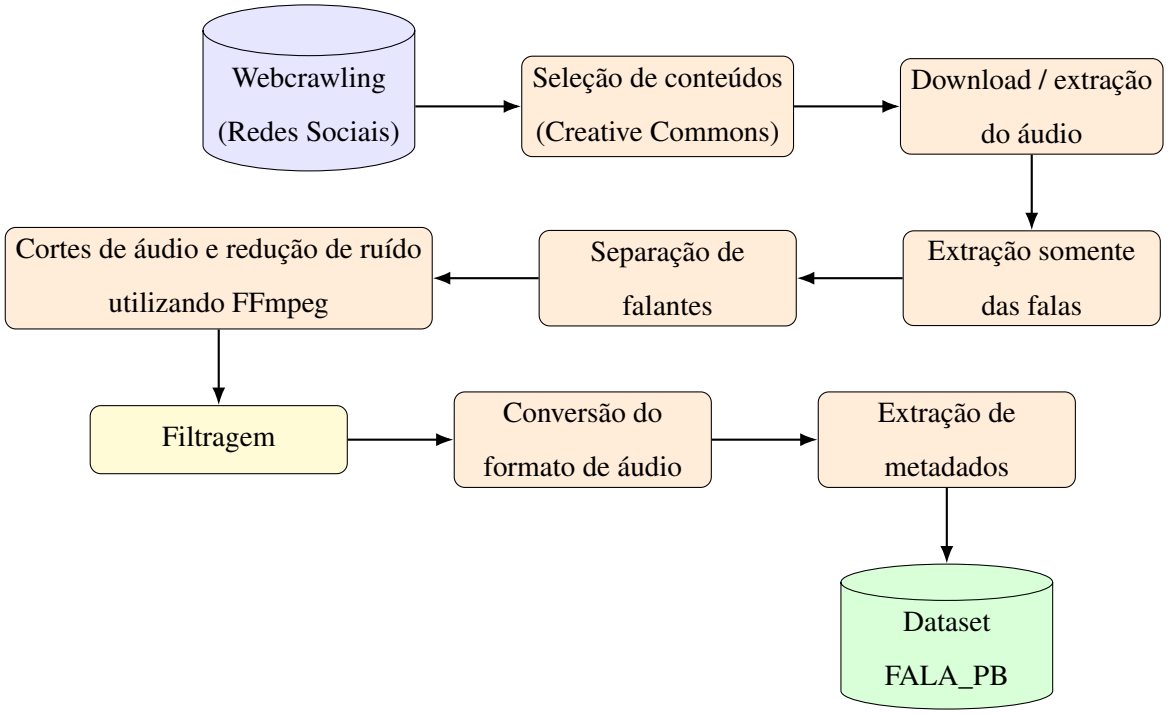
Como possibilidade de expansão futura da base FALA_PB, pretende-se investigar a utilização de técnicas de *web crawling* para coleta automatizada de amostras adicionais de fala em plataformas digitais e redes sociais. Essa abordagem busca ampliar a diversidade fonética,

prosódica e sociolinguística da base, permitindo a inclusão de diferentes perfis de falantes, contextos comunicativos e variações naturais da fala espontânea.

A proposta considera exclusivamente conteúdos disponibilizados sob licenças compatíveis com reutilização e modificação, como modalidades específicas de Creative Commons, respeitando os critérios legais e éticos relacionados ao uso de dados públicos. Além disso, futuras etapas deverão incluir mecanismos adicionais de filtragem automática, anonimização e validação da qualidade acústica dos materiais coletados.

Nesse contexto, pretende-se desenvolver pipelines automatizados capazes de realizar a extração, segmentação, transcrição e organização dos áudios coletados, reduzindo o esforço manual necessário na construção de grandes corpora regionais de fala. A Figura 6.1 apresenta uma proposta de fluxo para coleta e processamento de dados via *web crawling* em redes sociais e plataformas de compartilhamento de vídeo.

Figura 6.1: Fluxo de coleta via *webcrawling* em redes sociais e pré-processamento para extensão da base FALA_PB.



Fonte: elaborado pelo autor (2026).

6.2.3 Especialização Regional do Sotaque Paraibano

Outra possibilidade relevante para trabalhos futuros consiste na investigação mais aprofundada das diferentes variações regionais presentes dentro do próprio estado da Paraíba. Embora esta pesquisa tenha utilizado uma abordagem mais ampla voltada ao sotaque paraibano de maneira geral, observa-se que existem diferenças fonéticas, prosódicas e lexicais significativas entre distintas regiões do estado.

Dessa forma, futuras pesquisas podem investigar a construção de subconjuntos específicos de corpus com falas paraibanas organizadas por regiões, cidades ou outros perfis mais delimitados. Essa abordagem permitiria o treinamento de checkpoints mais especializados, potencialmente capazes de representar de maneira mais consistente características específicas de determinadas regiões da Paraíba.

Além disso, a especialização regional dos modelos pode contribuir para estudos relacionados à representatividade linguística, preservação de variações locais da fala e análise computacional de diferenças dialetais dentro do português brasileiro.

6.2.4 Arquiteturas com Modelagem Explícita de Sotaque

Uma direção complementar consiste em investigar arquiteturas que tratam o sotaque como fator explícito e separável da identidade do falante, em lugar de induzi-lo indiretamente por ajuste fino, como se fez nesta dissertação. Trabalhos como o NaturalSpeech 3 (Ju et al., 2024) e o CrossAccent-TTS (Annamdevula et al., 2026) decompõem a fala em fatores independentes e permitem controlar a intensidade do sotaque na síntese; procedimentos dedicados de conversão de sotaque (Liu et al., 2020) e estratégias de adaptação com dados limitados da variedade alvo (Zhou et al., 2023) seguem direção semelhante. A discussão da Seção 5.3.1 indica que é justamente essa separação explícita que falta aos modelos empregados aqui, nos quais a representação extraída do áudio de referência codifica essencialmente o timbre.

A aplicação dessas abordagens ao português brasileiro depende, contudo, de corpora multi-sotaque rotulados por variedade regional, cuja escassez foi discutida na Seção 3.3.3. A ampliação do *FALA_PB* e a construção de bases equivalentes para outras variedades do português brasileiro são, portanto, pré-requisito para essa linha de investigação, o que a articula diretamente às duas sugestões anteriores.

6.3 Considerações Finais

A incorporação de sotaques regionais em sistemas de síntese de fala é, antes de uma questão técnica, uma questão de representatividade. Os resultados deste trabalho oferecem evidências de que a diversidade linguística do português brasileiro pode ser contemplada em tecnologias de fala com recursos modestos e sem depender exclusivamente de plataformas proprietárias. Espera-se que os recursos e procedimentos desenvolvidos sirvam de ponto de partida para pesquisas futuras voltadas a outros sotaques e variedades regionais, contribuindo para sistemas de fala mais inclusivos e culturalmente representativos.

Referências Bibliográficas

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, 3 edition.

Amazon Web Services (2026). Amazon polly: Ai voice generator and text-to-speech. `https://aws.amazon.com/polly/`. Acesso em: 6 ago. 2026.

Annamdevula, R., Tatawat, A., Gudmalwar, A. P., Shah, N. J., e Wasnik, P. (2026). Crossaccent-tts: Cross-lingual accent-intensity controllable text-to-speech via disentangled speaker and accent representations. *arXiv preprint arXiv:2606.25403*.

Atamanenko, O., Chalova, A., Coombes, J., Cope, N., Dang, P., Deng, Z., Du, J., Ermolenko, M., et al. (2025). TTS-1 technical report. *arXiv preprint arXiv:2507.21138*. Relatório técnico do Inworld TTS-1, modelos TTS autorregressivos baseados em Transformer.

Bagno, M. (2007). *Nada na Língua é por Acaso: Por uma Pedagogia da Variação Linguística*. Parábola Editorial.

Batista, N. A. R. (2019). Estudo sobre identificação automática de sotaques regionais brasileiros baseada em modelagens estatísticas e técnicas de aprendizado de máquina. Dissertação de mestrado, Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP. Orientadores: Lee Luan Ling, Tiago Fernandes Tavares.

Berger, N., Sun, S., e Richmond, K. (2025). Non-standard accent TTS support via large multi-accent frontend pronunciation knowledge transfer. In *Proceedings of Interspeech 2025*, pages 2530–2534.

Cagliari, L. C. (1992). *Prosódia: algumas funções dos supra-segmentos*. Editora da UNICAMP.

Cambre, J. e Kulkarni, C. (2019). One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–19.

Canopy Labs (2025). Orpheus TTS. `https://github.com/canopyai/Orpheus-TTS`. Acesso em: 9 jul. 2026.

Casanova, E., Davis, K., Gölge, E., Gulea, I., et al. (2024). Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*. Disponível em arXiv.

Castilho, A. T. d. (2010). *Nova Gramática do Português Brasileiro*. Contexto.

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., e Chen, X. (2025). F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, Viena, Áustria. Association for Computational Linguistics.

Cui, W., Yu, D., Jiao, X., Meng, Z., Zhang, G., Wang, Q., Guo, S. Y., e King, I. (2025). Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970, Viena, Áustria. Association for Computational Linguistics.

de Morais, T. D. P. (2026). Incorporando Regionalismos e Sotaques em Modelos de Síntese de Fala - Corpus. Zenodo. Corpus disponível em: `https://zenodo.org/records/21251709`.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht.

ElevenLabs (2026). Elevenlabs: Ai voice generator & text to speech. `https://elevenlabs.io`. Acesso em: 9 jul. 2026.

Federico, M., Enyedi, R., Barra-Chicote, R., Giri, R., Isik, U., Krishnaswamy, A., e Sawaf, H. (2020). From speech-to-speech translation to automatic dubbing. In *Proc. 17th International Conference on Spoken Language Translation (IWSLT)*, pages 257–264.

Freitas, D. e Kouroupetroglou, G. (2008). Speech technologies for blind and low vision persons. *Technology and Disability*, 20(2):135–156.

Goodfellow, I., Bengio, Y., e Courville, A. (2016). *Deep Learning*. MIT Press.

Google Cloud (2026). Google cloud text-to-speech ai: Lifelike ai voices and speech synthesis. `https://cloud.google.com/text-to-speech`. Acesso em: 6 ago. 2026.

Han, D., Han, M., e Unsloth Team (2023). Unsloth. `https://github.com/unslothai/unsloth`. Biblioteca de otimização de ajuste fino de modelos de linguagem e fala.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1):81–88.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., e Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Hu, H., Zhu, X., He, T., Guo, D., Zhang, B., Wang, X., Guo, Z., Jiang, Z., Hao, H., Guo, Z., Zhang, X., Zhang, P., Yang, B., Xu, J., Zhou, J., e Lin, J. (2026). Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*.

Hunt, A. J. e Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 373–376. IEEE.

Hunt, M. J. (1990). Figures of merit for assessing connected-word recognisers. *Speech Communication*, 9(4):329–336.

ITU-T (1996). Methods for subjective determination of transmission quality. *ITU-T Recommendation P.800*.

Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., Wu, Z., Qin, T., Li, X.-Y., Ye, W., Zhang, S., Bian, J., He, L., Li, J., e Zhao, S. (2024). Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models.

Jurafsky, D. e Martin, J. H. (2009). *Speech and Language Processing*. Prentice Hall.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., e Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Klatt, D. H. (1987). Review of text-to-speech conversion for english. *The Journal of the Acoustical Society of America*, 82(3):737–793.

Kong, J., Kim, J., e Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *NeurIPS*.

Kong, Z., Ping, W., Huang, J., Zhao, K., e Catanzaro, B. (2021). Diffwave: A versatile diffusion model for audio synthesis. *ICLR*.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

Ladefoged, P. e Johnson, K. (2012). *A Course in Phonetics*. Cengage Learning.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Liao, S., Wang, Y., Li, T., Cheng, Y., Zhang, R., Zhou, R., e Xing, Y. (2024). Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*. Nesta dissertação foi utilizada a versão V1.5 do modelo, disponível em `https://github.com/fishaudio/fish-speech`.

Liu, S., Wang, D., Cao, Y., Sun, L., Wu, X., Kang, S., Wu, Z., Liu, X., Su, D., Yu, D., e Meng, H. (2020). End-to-end accent conversion without using native utterances. In *ICASSP 2020 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6289–6293.

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., e Wang, H.-M. (2019). MOSNet: Deep learning-based objective assessment for voice conversion. In *Proceedings of Interspeech 2019*, pages 1541–1545.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. *International Symposium on Music Information Retrieval*.

Marzal, A. e Vidal, E. (1993). Computation of normalized edit distance and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):926–932.

Microsoft (2026). Azure ai speech. `https://azure.microsoft.com/en-us/products/ai-foundry/tools/speech`. Acesso em: 6 ago. 2026.

Milan, G. G. (2021). gabriel-milan/sotaque-brasileiro: Sotaque Brasileiro.

Mittag, G., Naderi, B., Chehadi, A., e Möller, S. (2021). NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, pages 2127–2131.

Moraes, J. A. d. (2008). The pitch accents in brazilian portuguese: analysis by synthesis. *Proceedings of Speech Prosody*.

Morris, A. C., Maier, V., e Green, P. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768.

Oliveira, F. S. d., Gris, L. R. S., Ferreira, A. I. S., Rosa, A. S. d., Ferro Filho, A. C., Casanova, E., Shulby, C. D., Sousa, R. T., Silva, D. F. C., Soares, A. d. S., e Galvão Filho, A. R. (2026). Tagarela - a portuguese speech dataset from podcasts. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., e Kudinov, M. (2021). Grad-TTS: A diffusion probabilistic model for text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8599–8608.

Prenger, R., Valle, R., e Catanzaro, B. (2019). Waveglow: A flow-based generative network for speech synthesis. *ICASSP*.

Rabiner, L. e Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., e Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 28492–28518.

Ramos, G. (1938). *Vidas Secas*. Livraria José Olympio, Rio de Janeiro. Obra em domínio público no Brasil.

Reddy, C. K. A., Gopal, V., e Cutler, R. (2022). DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 886–890.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., e Liu, T.-Y. (2021). FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations (ICLR)*. Disponível também como arXiv:2006.04558.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., e Liu, T.-Y. (2019). FastSpeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., et al. (2018). Natural tts synthesis by conditioning WaveNet on Mel spectrogram predictions. In *ICASSP*, pages 4779–4783. IEEE.

Silva, T. C. (2011). *Fonética e Fonologia do Português: Roteiro de Estudos e Guia de Exercícios*. Contexto.

Smith, S. W. (1997). *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, San Diego, CA.

Stein, C. C. et al. (2013). Corpus linguajar paraibano (colingpb). `http://cchla.ufpb.br/colingpb`. Disponível em: `http://cchla.ufpb.br/colingpb`.

Stevens, S. S., Volkmann, J., e Newman, E. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

Tan, X., Qin, T., Soong, F., e Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.

Tinchev, G., Czarnowska, M., Deja, K., Yanagisawa, K., e Cotescu, M. (2023). Modelling low-resource accents without accent-specific TTS frontend. In *ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., e Oura, K. (2013). Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., e Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

van den Oord, A., Vinyals, O., e Kavukcuoglu, K. (2017). Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., e Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wagner, R. A. e Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Xie, T., Rong, Y., Zhang, P., Wang, W., e Liu, L. (2025). Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 764–791, Suzhou, China. Association for Computational Linguistics. Disponível também como arXiv:2412.06602.

Yang, Y., Liu, S., Li, J., et al. (2025). Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis. *Proceedings*. arXiv:2504.10352.

Zen, H., Tokuda, K., e Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

Zezario, R. E., Fu, S.-W., Chen, F., Fuh, C.-S., Wang, H.-M., e Tsao, Y. (2023). Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:54–70.

Zhang, W., Yeh, C.-C., Beckman, W., Raitio, T., Rasipuram, R., Golipour, L., e Winarsky, D. (2023). Audiobook synthesis with long-form neural text-to-speech. In *Proc. 12th ISCA Speech Synthesis Workshop (SSW)*, pages 139–143.

Zhou, X. et al. (2023). Accented text-to-speech synthesis with limited data. *arXiv preprint arXiv:2305.04816*.

Apêndice A

Prompt para Geração das Mensagens de Coleta Colaborativa

Este apêndice reproduz o *prompt* utilizado para gerar, via ChatGPT (modelo GPT-4o), as mensagens de solicitação das três frases de leitura enviadas aos participantes da coleta colaborativa (Seção 4.4.2).

> Gere uma mensagem curta para WhatsApp pedindo participação em uma coleta de áudio acadêmica para dissertação de mestrado sobre síntese de voz e sotaque paraibano.
>
> A mensagem deve ser curta, direta e informal, evitando verbosidade.
>
> Estrutura obrigatória:
>
> 1. Pedido para a pessoa enviar 3 áudios separados no WhatsApp; 2. Explicação curta de que os áudios serão usados apenas para fins acadêmicos da dissertação; 3. Informar que não haverá identificação dos participantes nem uso comercial; 4. Exibir 3 frases numeradas.
>
> As frases devem seguir obrigatoriamente estas regras:
>
> • Cada frase deve durar aproximadamente 10 segundos de leitura em velocidade normal; • As frases devem soar naturais, espontâneas e cotidianas; • Não usar frases excessivamente poéticas, técnicas, formais ou artificiais; • Variar estruturas fonéticas, construções gramaticais e vocabulário; • Evitar frases muito curtas; • Cada frase deve ser independente das demais; • Variar fortemente os temas entre gerações; • Evitar repetir contextos recorrentes como trânsito, chuva, feira, ônibus, trabalho, praça, diretoria, escola ou caminhada; • Buscar diversidade de situações do cotidiano; • Misturar ações, descrições, opiniões, lembranças, tarefas, lazer, tecnologia, família, comida, compras, jogos, internet, estudos, pets, exercícios, viagens, eventos, séries, música, rotina doméstica e conversas informais; • Evitar que múltiplas frases da mesma geração compartilhem o mesmo ambiente ou assunto central.
>
> Restrições obrigatórias:
>
> • Não mencionar explicitamente que existem restrições fonéticas; • Não explicar as regras no resultado final; • Entregar a resposta já pronta para copiar e enviar.
>
> Evite gerar frases semanticamente parecidas entre diferentes execuções. Priorize variedade temática e lexical.

Exemplos de temas aceitáveis: configurar computador novo; preparar comida em casa; jogar videogame online; assistir filme ou série; organizar quarto ou documentos; conversar com amigos; passear com cachorro; fazer compras em mercado; resolver algo pelo celular; academia ou exercício físico; planejamento de viagem; problemas com internet; eventos familiares; rotina universitária; música, instrumentos ou shows; uso de aplicativos; situações engraçadas do cotidiano; atividades de fim de semana.

Exemplos variados de frases aceitáveis:

"Depois que troquei de celular, ainda estou tentando organizar todos os aplicativos e recuperar algumas fotos antigas que estavam salvas."

"Enquanto estava preparando o jantar ontem à noite, fiquei ouvindo música e respondendo mensagens que tinham chegado durante o dia."

"O diretor comentou que a atividade de terça-feira vai precisar de confirmação antecipada por causa da quantidade de participantes inscritos."

"Meu cachorro ficou correndo pela casa inteira depois que percebeu que alguém tinha pegado a coleira para sair."

"Passei boa parte da manhã tentando descobrir por que a televisão não estava conectando direito na internet do apartamento."

"Durante a viagem do fim de semana, a gente decidiu parar em uma cidade pequena para experimentar a comida local e descansar um pouco."

A.1 Mensagens Enviadas aos Participantes

A mensagem de convite enviada após a apresentação seguiu o seguinte texto:

> Preciso fazer uma pequena coleta de falas para uma pesquisa acadêmica. A ideia é ajudar na construção de um corpus linguístico voltado ao sotaque da região. Seriam 3 frases que você poderia mandar aqui pelo WhatsApp lendo em voz alta, de preferência em um ambiente silencioso.
>
> Posso te mandar as frases?

Um exemplo de mensagem efetivamente enviada a um participante que aceitou o convite é:

> Você poderia me mandar 3 áudios separados no WhatsApp, um para cada frase abaixo?
>
> Os áudios serão usados somente para fins acadêmicos da minha dissertação, sem identificação dos participantes e sem qualquer uso comercial.
>
> Frase 1: "Esses dias eu resolvi tentar aprender a fazer café de um jeito diferente e fiquei surpreso com a quantidade de detalhes que existem nisso."

Frase 2: “Enquanto estava organizando os arquivos antigos do notebook durante a madrugada, fiquei encontrando fotos e vídeos que eu nem lembrava mais que existiam.”

Frase 3: “A atendente explicou que algumas atividades do aplicativo precisariam de atualização antes de voltar a funcionar normalmente no celular.”

Apêndice B

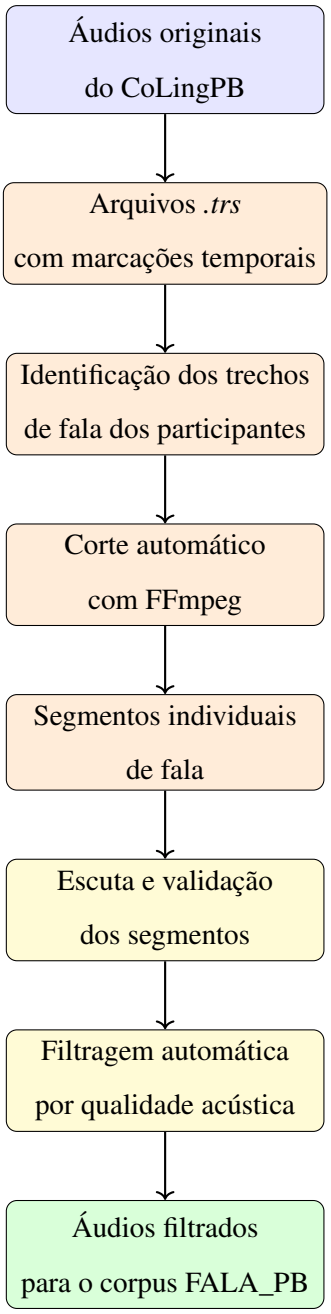
Processamento do Corpus Linguístico da Paraíba (CoLingPB)

Este apêndice detalha o processamento aplicado ao CoLingPB (Stein et al., 2013) para adequá-lo ao treino dos modelos de síntese de fala, complementando o resumo apresentado na Seção 4.4.

B.1 Extração e Segmentação

O CoLingPB é disponibilizado como entrevistas completas, acompanhadas de arquivos com extensão *.trs*, contendo marcações temporais e identificação do falante. A partir dessas marcações, os trechos correspondentes à fala dos entrevistados foram recortados automaticamente com a ferramenta FFmpeg, transformando cada entrevista em segmentos individuais de fala. Em seguida, aplicou-se a triagem descrita na Seção 4.4, que combina escuta manual e o preditor *DNSMOS* (Reddy et al., 2022), com corte pelo escore *ovrl*, para remover trechos com ruído excessivo ou baixa inteligibilidade. A Figura B.1 ilustra esse fluxo.

Figura B.1: Fluxo de extração, segmentação e filtragem dos áudios do CoLingPB.



Fonte: elaborado pelo autor (2026).

B.2 Codificação dos Arquivos Originais

No corpus original CoLingPB (Stein et al., 2013), cada gravação de entrevista segue um padrão no nome do arquivo *.wav*, em que trechos do identificador indicam *onde* e *quem* foi gravado. O formato geral é:

`brPB{NN}_g{idade}{opcionais}{sexo}{informante}.wav`

em que:

- `brPB`: Brasil, estado da Paraíba (corpus CoLingPB);
- `NN` (dois dígitos): código do município (`01` = João Pessoa, `07` = Campina Grande, `25` = Cajazeiras, entre outros; 26 municípios mapeados na Tabela 5.1);
- `g` + número: faixa etária: `g1` adolescente (14–17 anos), `g2` adulto (25–45), `g3` idoso (60+);
- `u` (opcional, junto ao `g`): informante graduado (`g2u` ou `g3u`);
- `a` ou `b` (opcional): zona urbana ou rural;
- `F` ou `M`: sexo do participante;
- dois dígitos finais: número do informante naquele município/grupo.

A Tabela B.1 resume essas partes do padrão de nomenclatura do corpus original (Stein et al., 2013).

Tabela B.1: Partes do padrão de nomenclatura dos arquivos originais do CoLingPB.

Parte	**Significado**
`brPB`	Brasil, estado da Paraíba (corpus CoLingPB)
`01`–`26`	Código do município (dois dígitos; mapa em 26 localidades, vide Tabela 5.1)
`g1`, `g2`, `g3`	Faixa etária: adolescente (14–17), adulto (25–45), idoso (60+)
`u` (opcional)	Informante graduado (`g2u`, `g3u`)
`a` / `b` (opcional)	Zona urbana / rural
`F` / `M`	Sexo: feminino / masculino
dois dígitos finais	Número do informante no município/grupo (ex.: `01`)

Fonte: elaborada pelo autor (2026), com base em (Stein et al., 2013).

Por exemplo, `brPB01_g1aF01.wav` corresponde a João Pessoa, adolescente, zona urbana, feminino, informante 01. Como o sufixo sociolinguístico admite variações de or-

dem e de separadores (por exemplo, `g3au` ou `g3ua`), os campos foram extraídos por um decodificador que normaliza essas variantes antes do cruzamento com os metadados.

O rótulo `g4` não consta como faixa oficial na documentação do corpus; após consulta ao professor Cirineu Cecote Stein (responsável pelo CoLingPB), adotou-se neste trabalho o tratamento de `g4` como equivalente a idoso graduado (`g3u`). Há ainda um caso pontual (`brPB25_g4aF01`) com regras específicas de município/nome, tratado de forma explícita no decodificador.

B.3 Nomenclatura dos Segmentos Processados

A partir das entrevistas originais, cada corte de fala do entrevistado recebeu um nome que preserva a gravação-fonte e explicita o intervalo temporal extraído do arquivo *.trs* disponibilizado pelo CoLingPB. O padrão adotado utiliza dois sublinhados consecutivos (__) como separador entre campos:

`{gravação_original}__{falante}__{início_ms}-{fim_ms}.wav`

em que:

- `gravação_original`: nome do arquivo *.wav* da entrevista (ex.: `brPB01_g1aF01`), mantendo a codificação da Tabela B.1;
- `falante`: identificador do falante no *.trs* (por exemplo, código do entrevistado em oposição ao entrevistador);
- `início_ms` e `fim_ms`: instantes de início e fim do trecho, em milissegundos, lidos das marcações temporais do *.trs* e utilizados pelo FFmpeg na segmentação.

Exemplo ilustrativo de segmento gerado:

`brPB01_g1aF01__participante__125000-142500.wav`

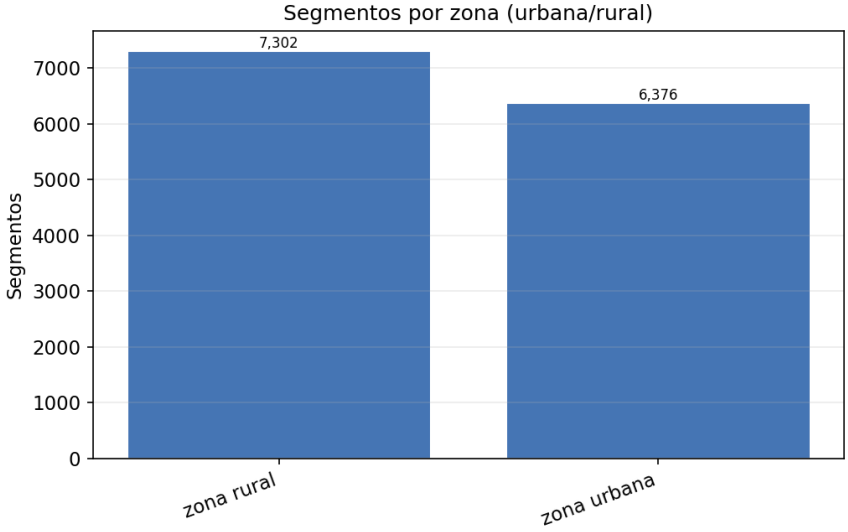
Esse esquema garante rastreabilidade de cada um dos 33.649 segmentos até a entrevista original e aos metadados sociolinguísticos decodificados a partir do prefixo `brPB`, em alinhamento com um arquivo de metadados no formato *.csv* verificado no disco: os 33.649 segmentos presentes no disco correspondem, um a um, às linhas do arquivo de metadados, sem sobras de nenhum dos dois lados.

B.4 Caracterização dos Segmentos do CoLingPB após Filtragem

Os gráficos das Figuras B.2 a B.6 descrevem o conjunto de 33.649 segmentos do CoLingPB após as etapas de segmentação e filtragem de qualidade acústica (Seção 4.4). Em todos os casos, o eixo ou as barras representam contagem de segmentos (*.wav*), e não o número de entrevistas originais (324).

A Figura B.2 apresenta a distribuição dos segmentos filtrados por zona do estado da Paraíba (codificação regional do corpus).

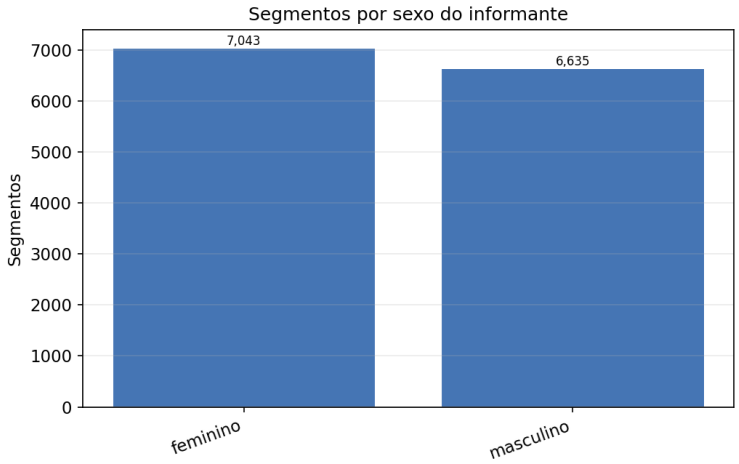
Figura B.2: Distribuição dos segmentos do CoLingPB processado por zona após filtragem.



Fonte: elaborado pelo autor (2026), com dados de (Stein et al., 2013).

A Figura B.3 mostra a distribuição por sexo do participante, conforme codificação das entrevistas originais.

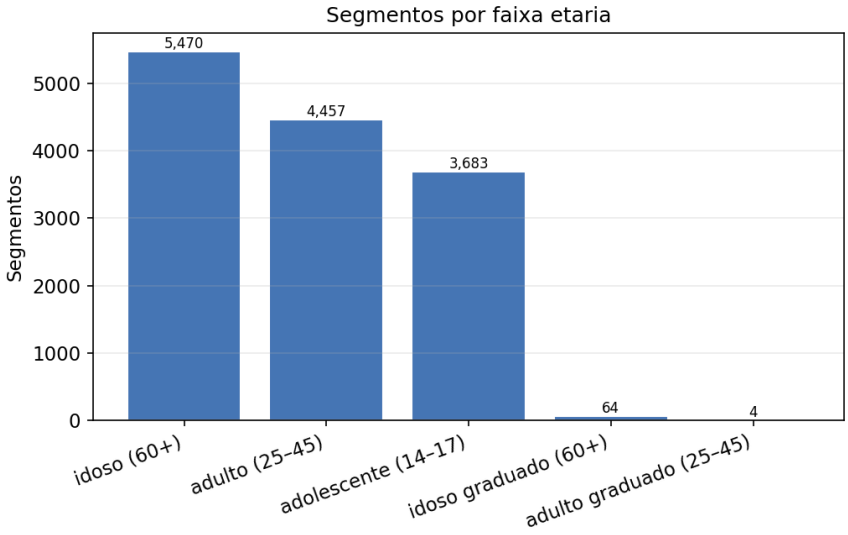
Figura B.3: Distribuição dos segmentos do CoLingPB processado por sexo do participante após filtragem.



Fonte: elaborado pelo autor (2026), com dados de (Stein et al., 2013).

A Figura B.4 apresenta a distribuição por faixa etária.

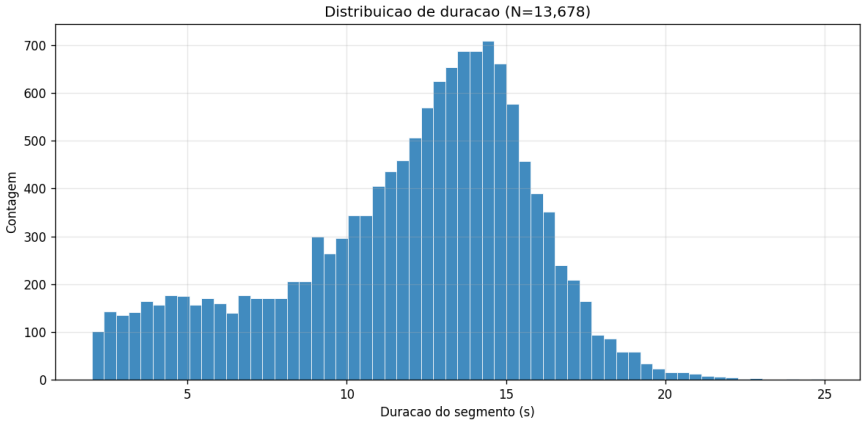
Figura B.4: Distribuição dos segmentos do CoLingPB processado por faixa etária após filtragem.



Fonte: elaborado pelo autor (2026), com dados de (Stein et al., 2013).

A Figura B.5 exibe o histograma da duração em segundos de cada segmento de áudio filtrado.

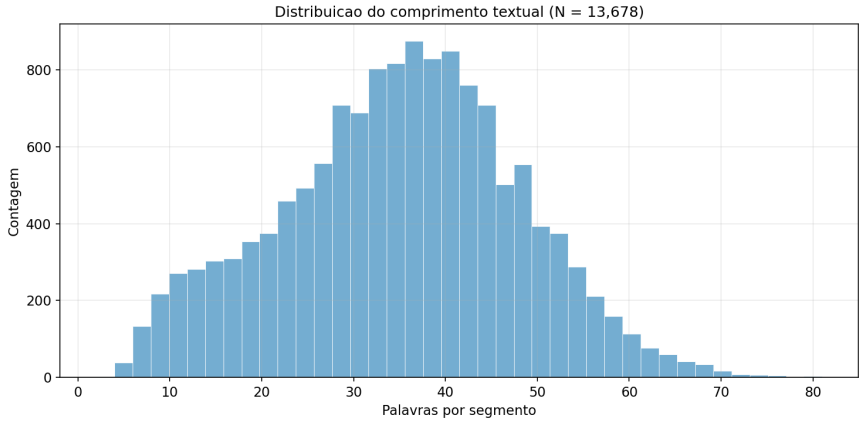
Figura B.5: Histograma da duração (em segundos) dos segmentos do CoLingPB processado (após filtragem).



Fonte: elaborado pelo autor (2026), com dados de (Stein et al., 2013).

Por fim, a Figura B.6 apresenta a distribuição do número de palavras por segmento, derivada das transcrições associadas a cada corte.

Figura B.6: Distribuição do número de palavras por segmento no CoLingPB processado após filtragem.



Fonte: elaborado pelo autor (2026), com dados de (Stein et al., 2013).

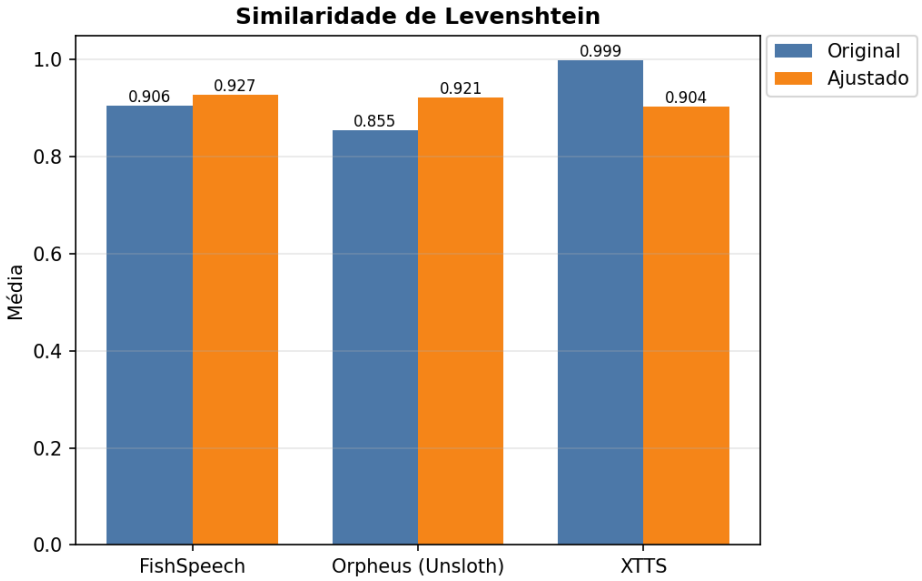
Apêndice C

Conjunto Completo de Gráficos da Avaliação Automática

Este apêndice reúne o conjunto completo de gráficos gerados na avaliação automática (Seção 5.4), para todas as métricas calculadas. Para cada métrica apresentam-se três visualizações complementares, todas calculadas sobre as 100 sentenças de teste em cada modelo e condição (ver Seção 5.4.1): o gráfico de barras das médias por modelo e condição, o diagrama de caixa (*boxplot*) da distribuição por modelo e condição e o histograma das frequências por modelo e condição. Em todos eles, a cor azul representa a condição original (modelo base) e a cor laranja, a condição ajustada (com sotaque paraibano); independentemente da cor, as duas condições também aparecem identificadas como *original* e *ajustado* no eixo horizontal (boxplots e histogramas) ou na legenda (gráficos de barras), permitindo a leitura em impressão em preto e branco.

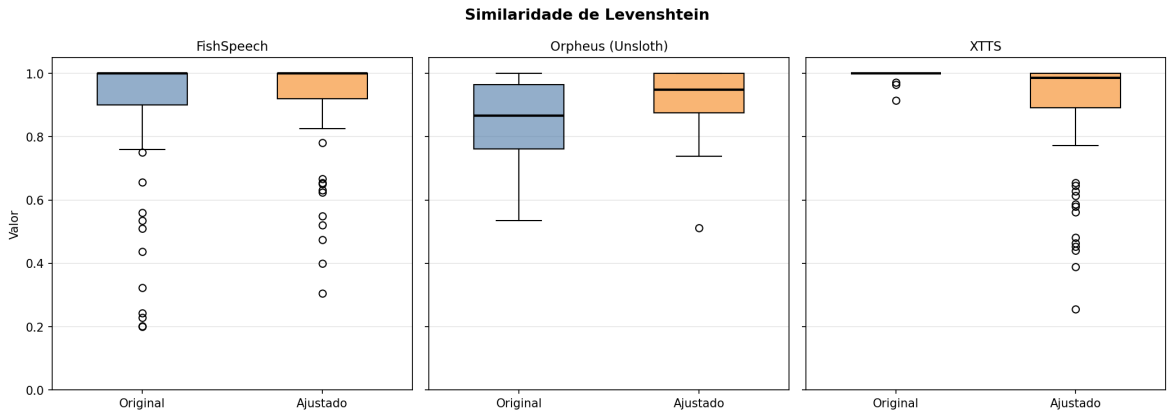
C.1 Inteligibilidade: Similaridade de Levenshtein

Figura C.1: Similaridade de Levenshtein: médias por modelo e condição.



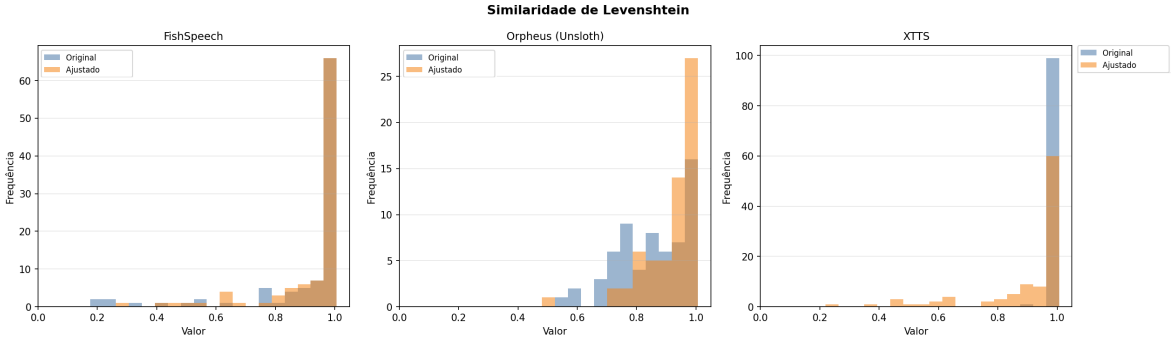
Fonte: elaborado pelo autor (2026).

Figura C.2: Similaridade de Levenshtein: distribuição por modelo e condição.



Fonte: elaborado pelo autor (2026).

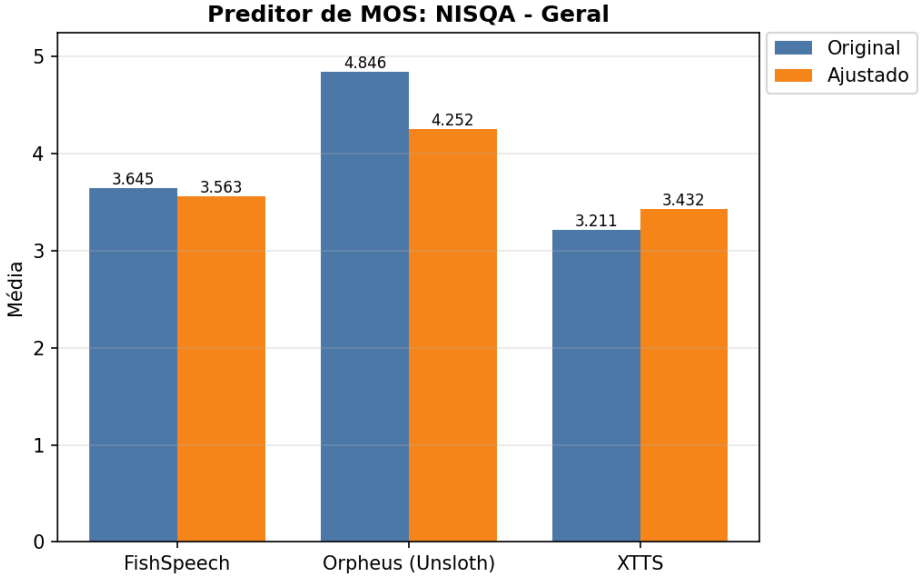
Figura C.3: Similaridade de Levenshtein: histograma por modelo e condição.



Fonte: elaborado pelo autor (2026).

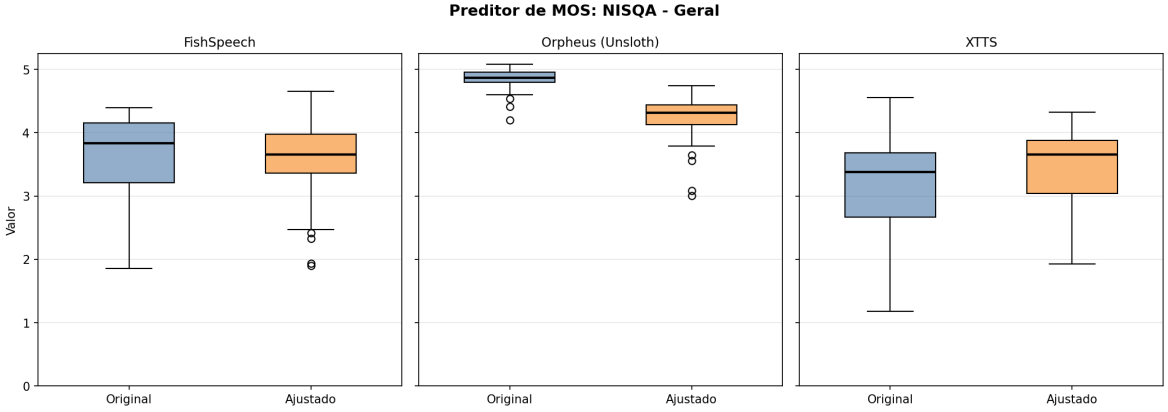
C.2 MOS Principal: *NISQA overall*

Figura C.4: *NISQA overall* (MOS principal): médias por modelo e condição.



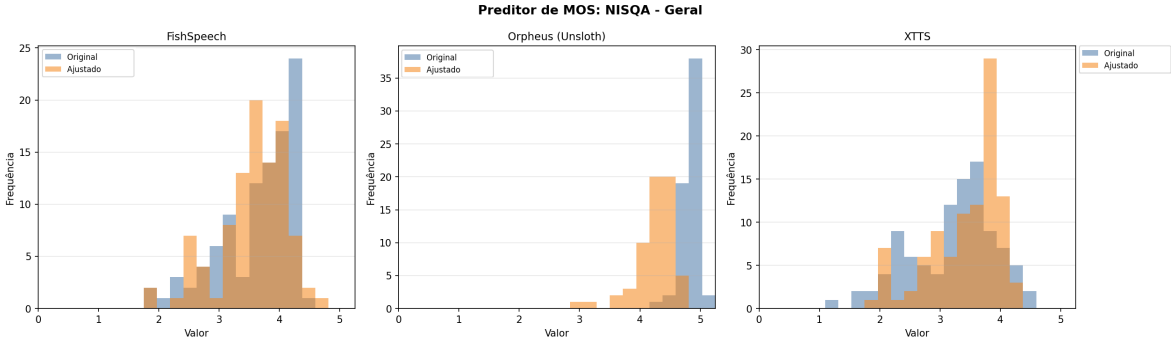
Fonte: elaborado pelo autor (2026).

Figura C.5: *NISQA overall*: distribuição por modelo e condição.



Fonte: elaborado pelo autor (2026).

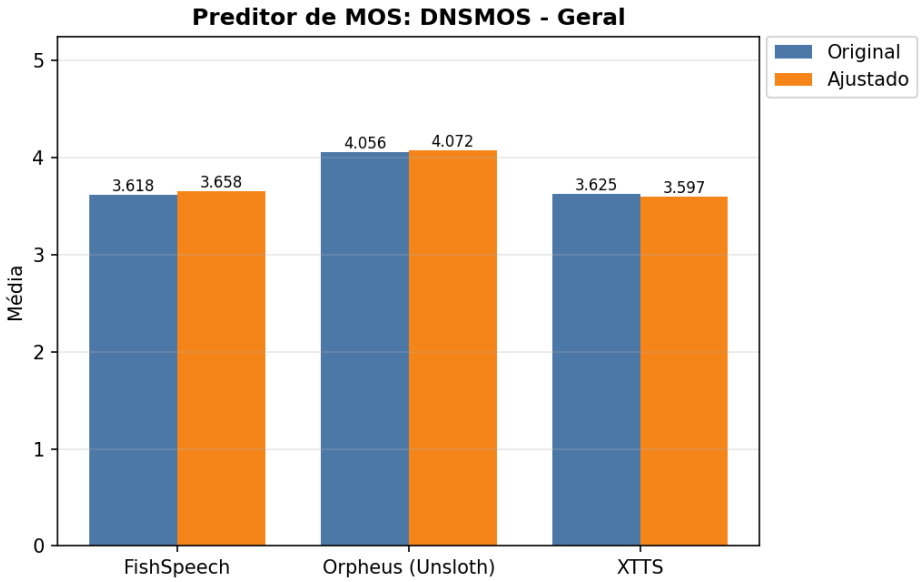
Figura C.6: *NISQA overall*: histograma por modelo e condição.



Fonte: elaborado pelo autor (2026).

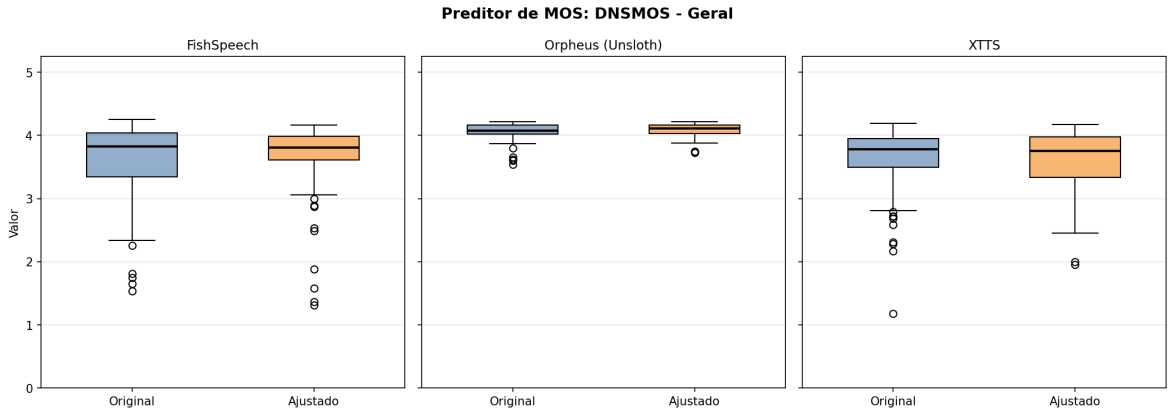
C.3 *DNSMOS*: Qualidade Global (*overall*)

Figura C.7: *DNSMOS overall*: médias por modelo e condição.



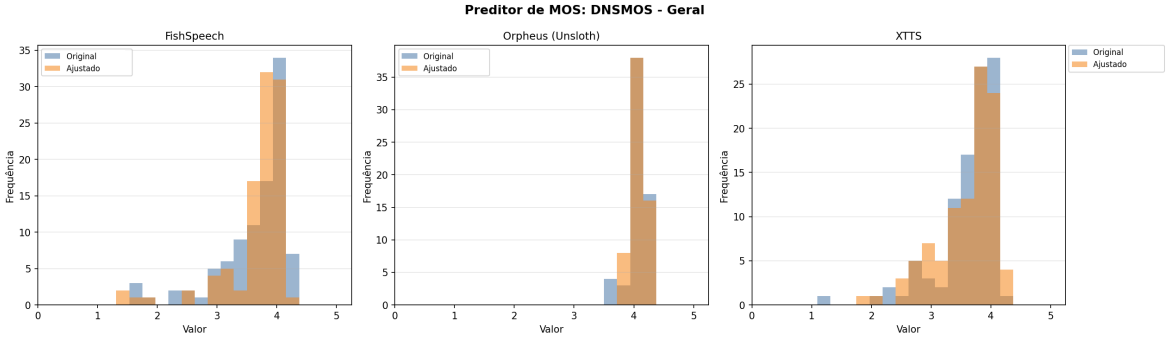
Fonte: elaborado pelo autor (2026).

Figura C.8: *DNSMOS overall*: distribuição por modelo e condição.



Fonte: elaborado pelo autor (2026).

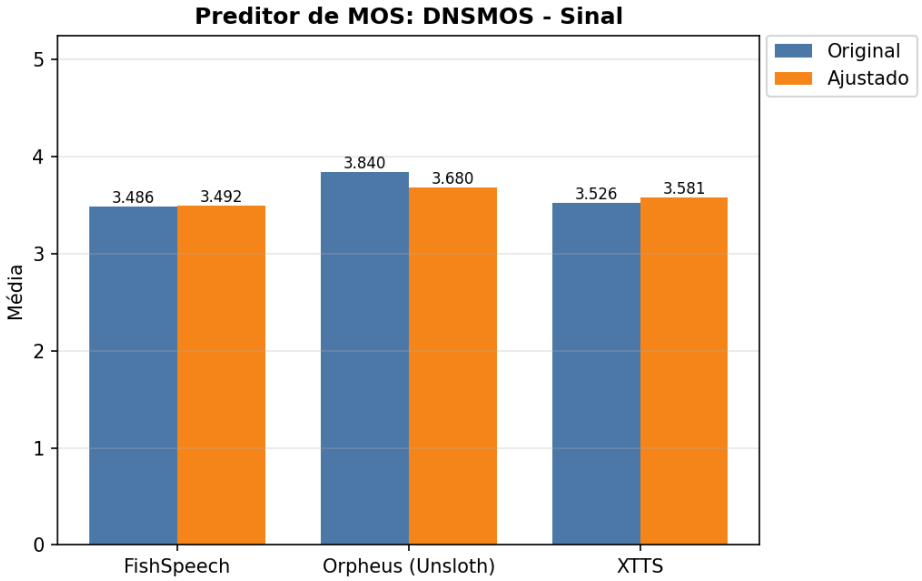
Figura C.9: *DNSMOS overall*: histograma por modelo e condição.



Fonte: elaborado pelo autor (2026).

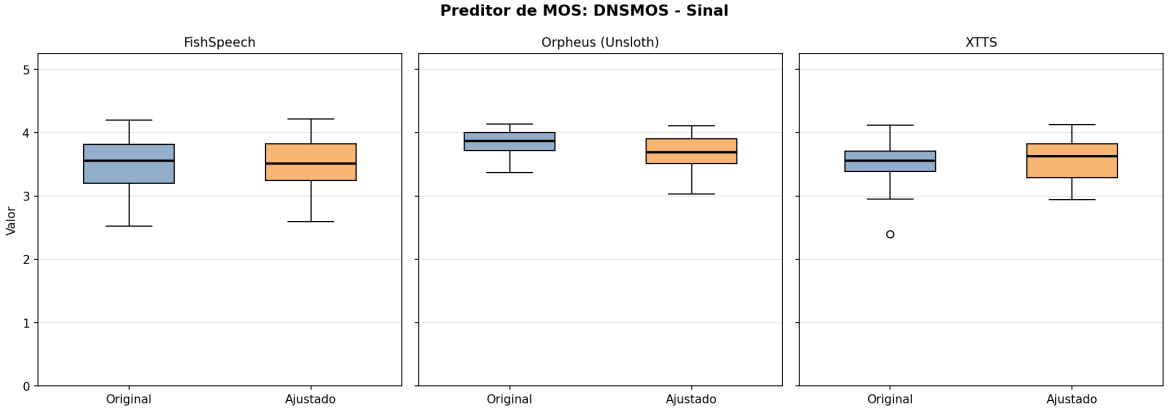
C.4 *DNSMOS*: Qualidade do Sinal de Voz (*sig*)

Figura C.10: *DNSMOS sig*: médias por modelo e condição.



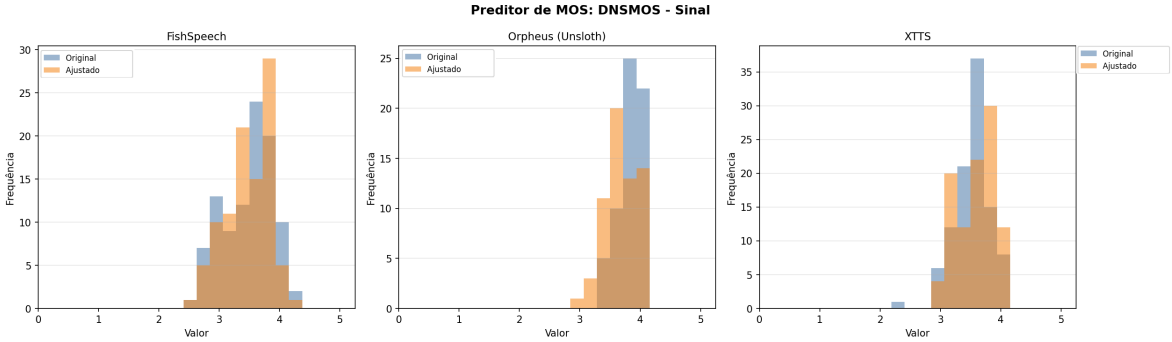
Fonte: elaborado pelo autor (2026).

Figura C.11: *DNSMOS sig*: distribuição por modelo e condição.



Fonte: elaborado pelo autor (2026).

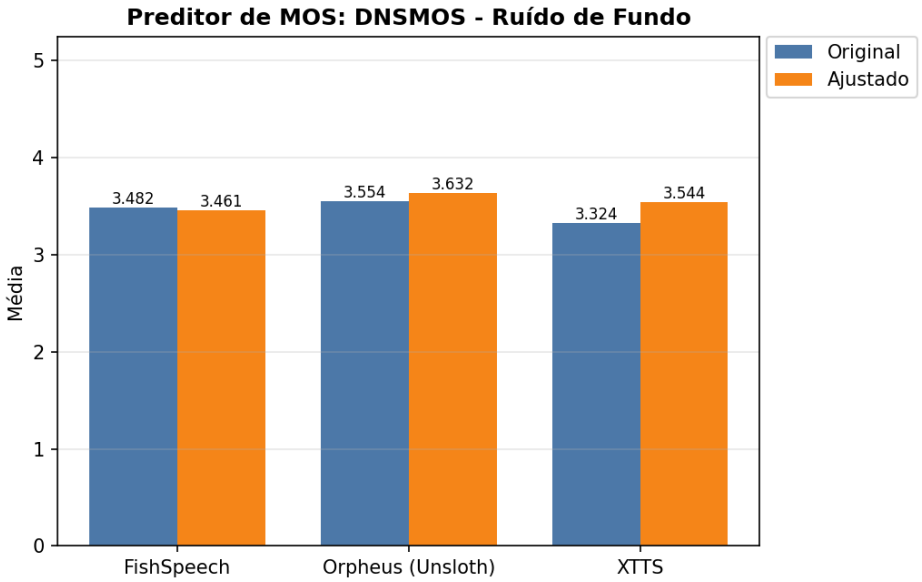
Figura C.12: *DNSMOS sig*: histograma por modelo e condição.



Fonte: elaborado pelo autor (2026).

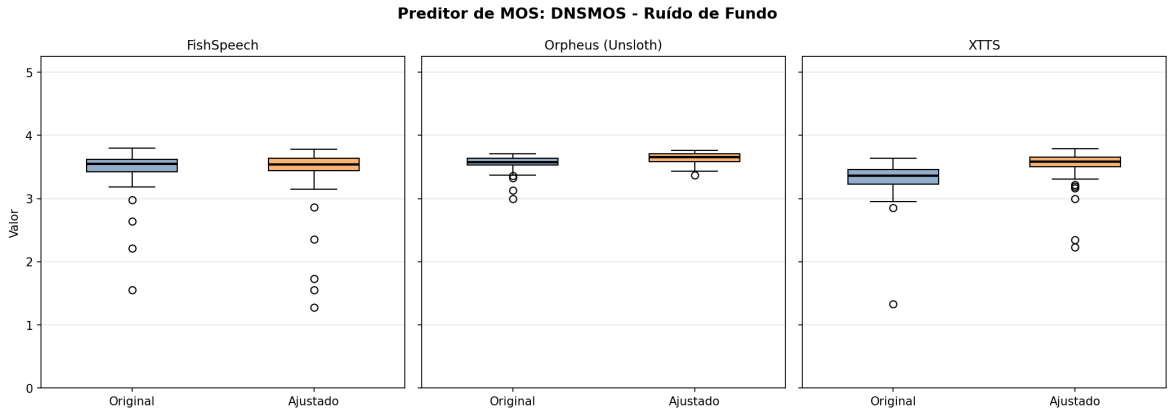
C.5 *DNSMOS*: Qualidade do Fundo (*bak*)

Figura C.13: *DNSMOS bak*: médias por modelo e condição.



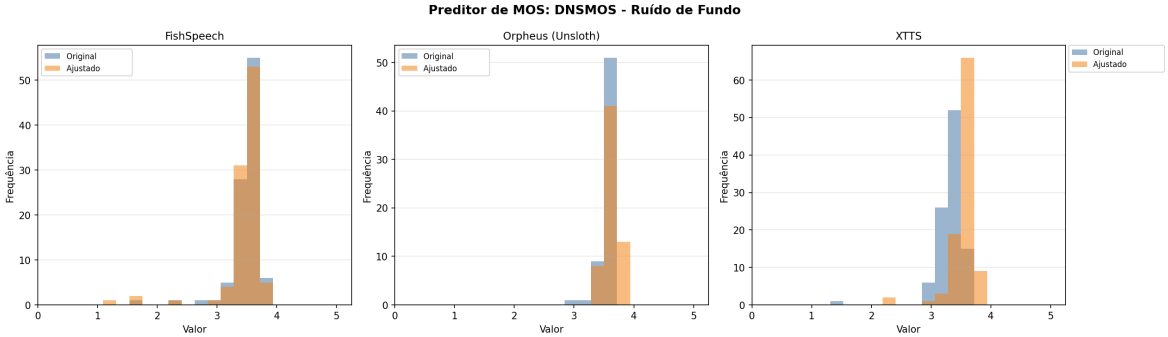
Fonte: elaborado pelo autor (2026).

Figura C.14: *DNSMOS bak*: distribuição por modelo e condição.



Fonte: elaborado pelo autor (2026).

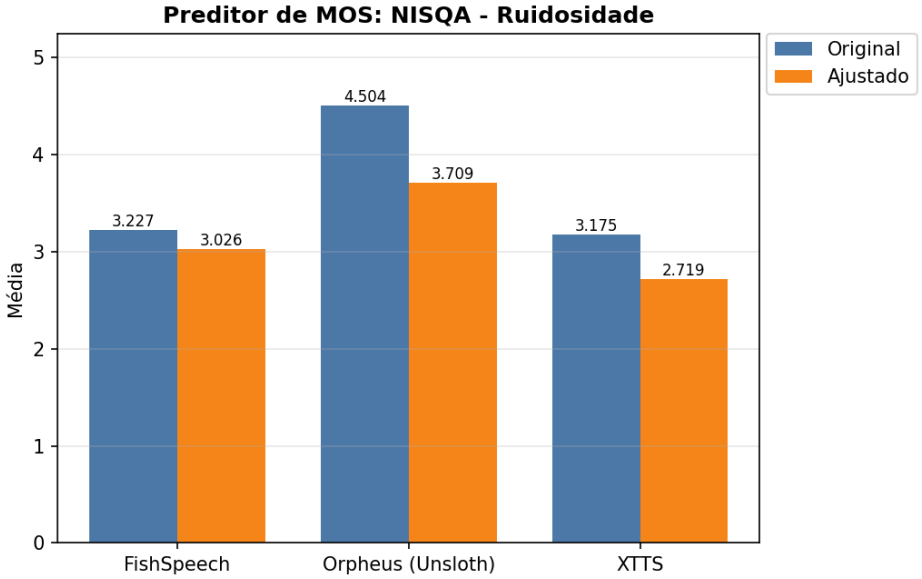
Figura C.15: *DNSMOS bak*: histograma por modelo e condição.



Fonte: elaborado pelo autor (2026).

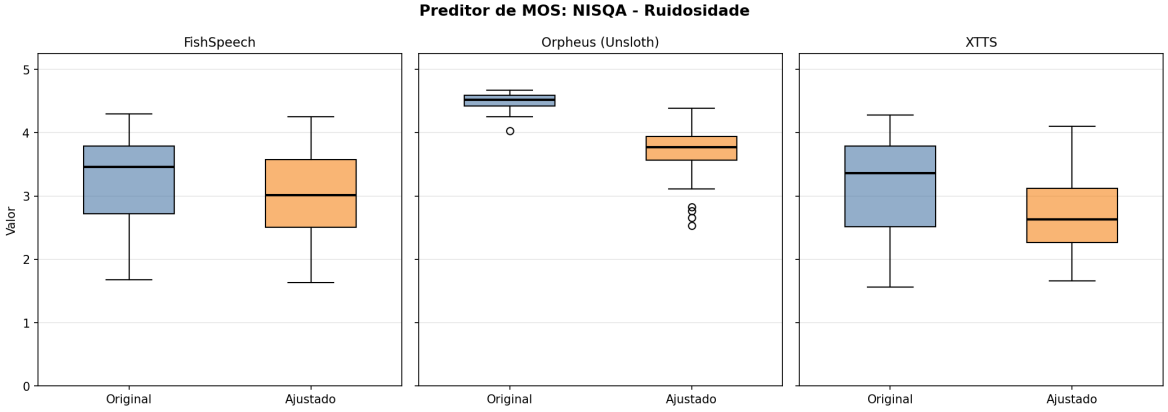
C.6 *NISQA*: Ruído Percebido (*noisiness*)

Figura C.16: *NISQA noisiness*: médias por modelo e condição.



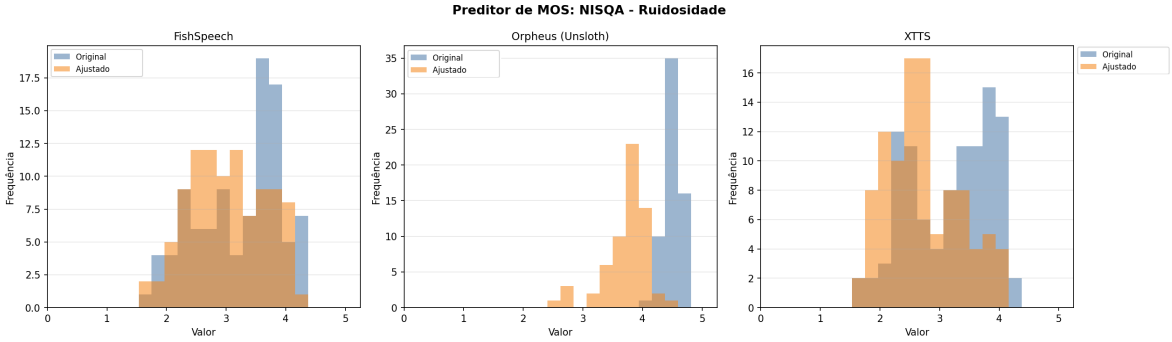
Fonte: elaborado pelo autor (2026).

Figura C.17: *NISQA noisiness*: distribuição por modelo e condição.



Fonte: elaborado pelo autor (2026).

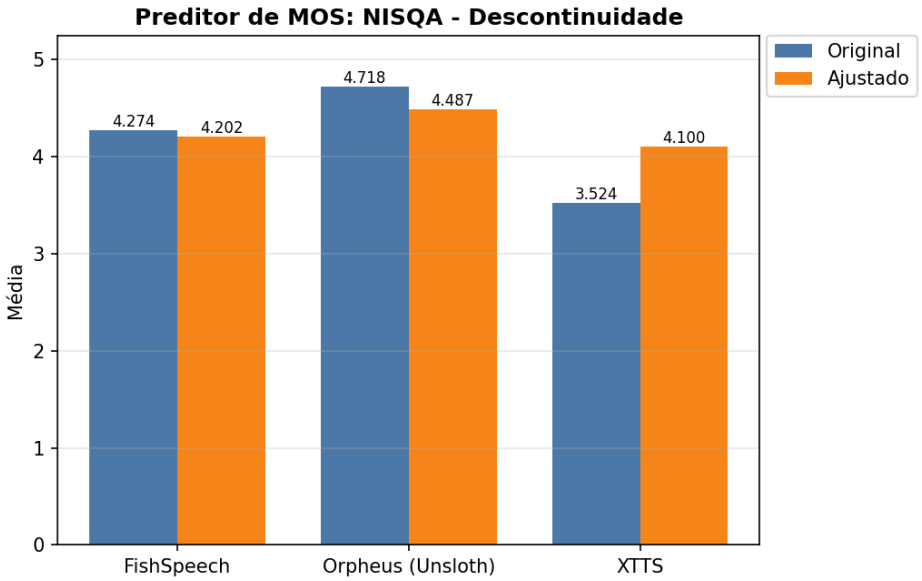
Figura C.18: *NISQA noisiness*: histograma por modelo e condição.



Fonte: elaborado pelo autor (2026).

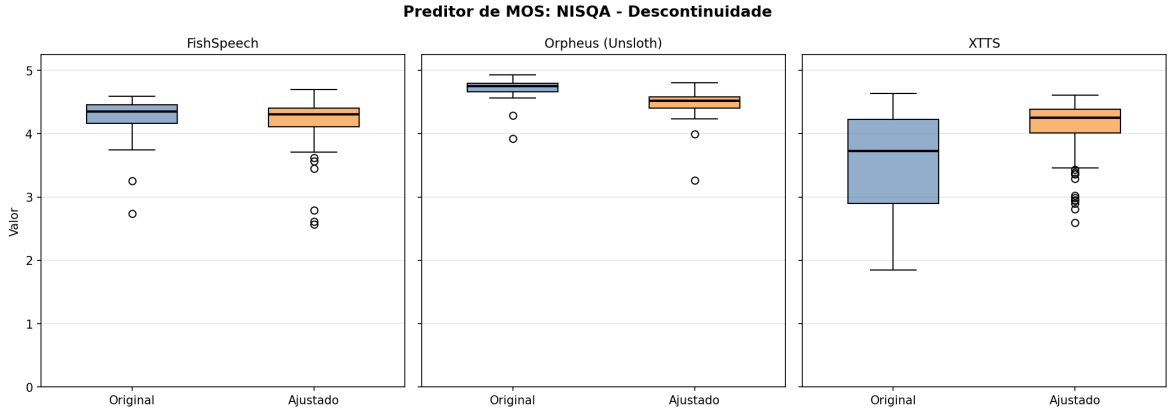
C.7 *NISQA*: Descontinuidade (*discontinuity*)

Figura C.19: *NISQA discontinuity*: médias por modelo e condição.



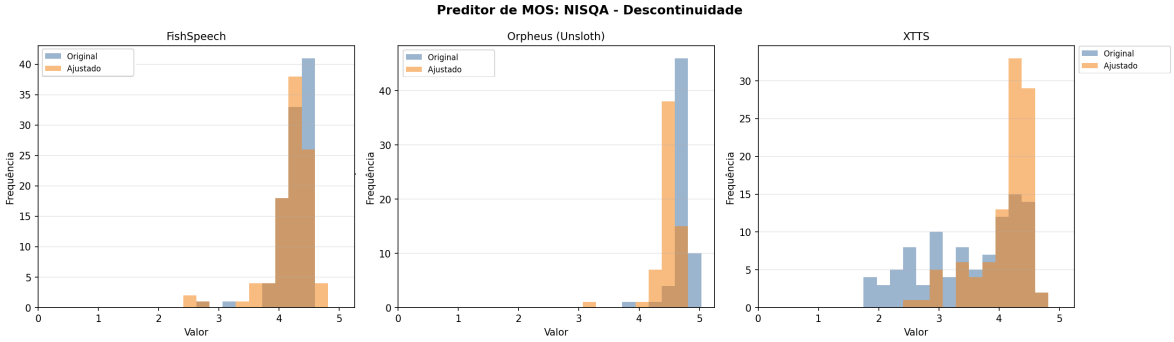
Fonte: elaborado pelo autor (2026).

Figura C.20: *NISQA discontinuity*: distribuição por modelo e condição.



Fonte: elaborado pelo autor (2026).

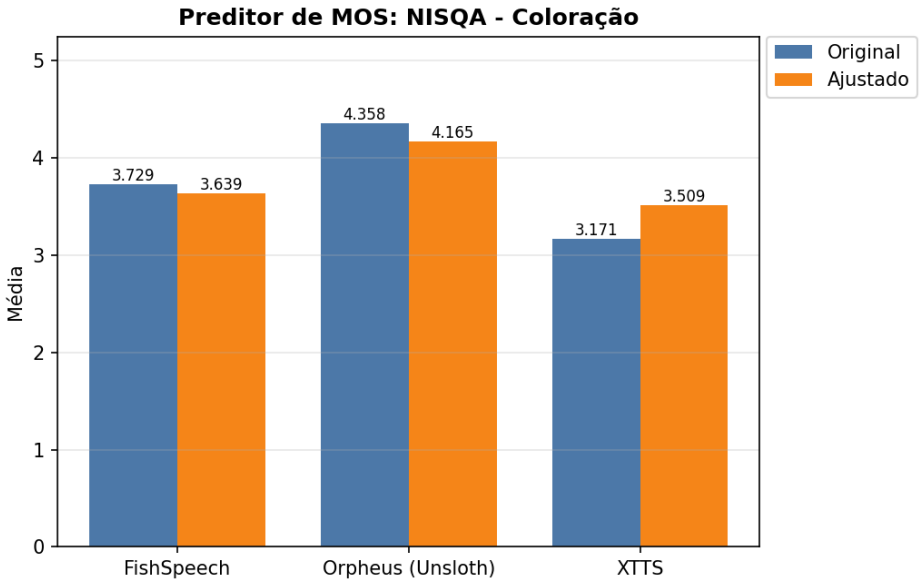
Figura C.21: *NISQA discontinuity*: histograma por modelo e condição.



Fonte: elaborado pelo autor (2026).

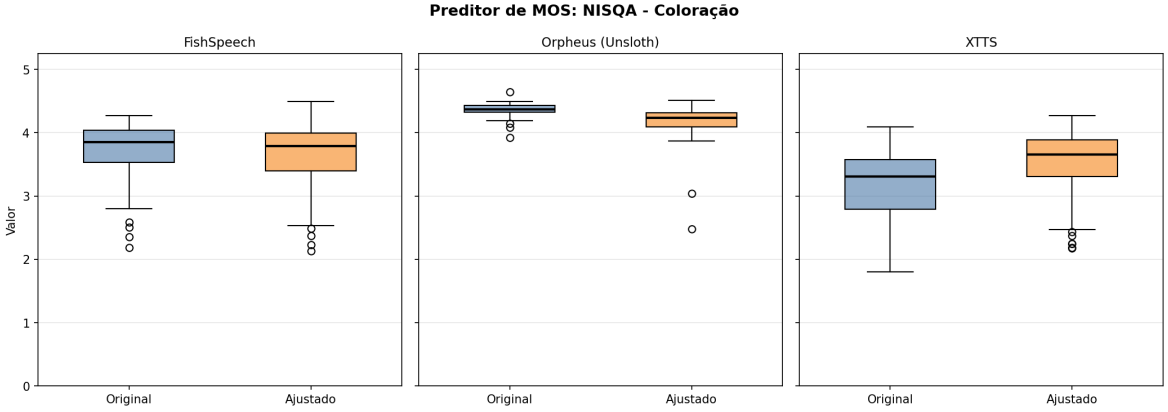
C.8 *NISQA*: Coloração (*coloration*)

Figura C.22: *NISQA coloration*: médias por modelo e condição.



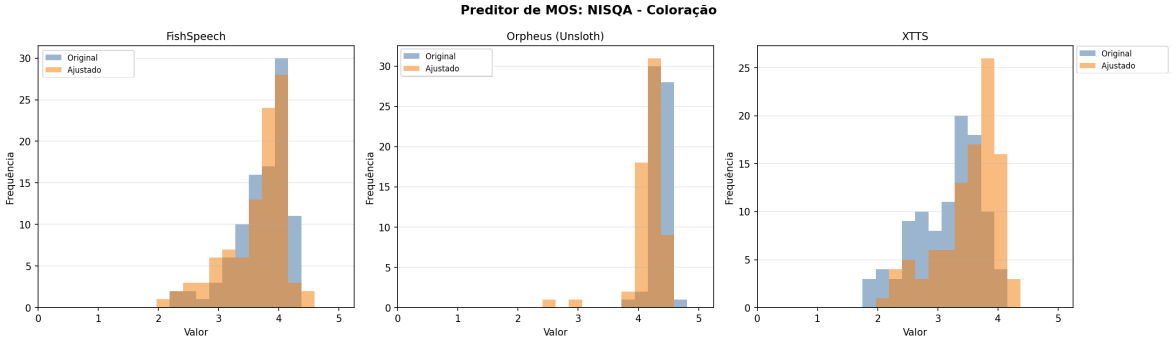
Fonte: elaborado pelo autor (2026).

Figura C.23: *NISQA coloration*: distribuição por modelo e condição.



Fonte: elaborado pelo autor (2026).

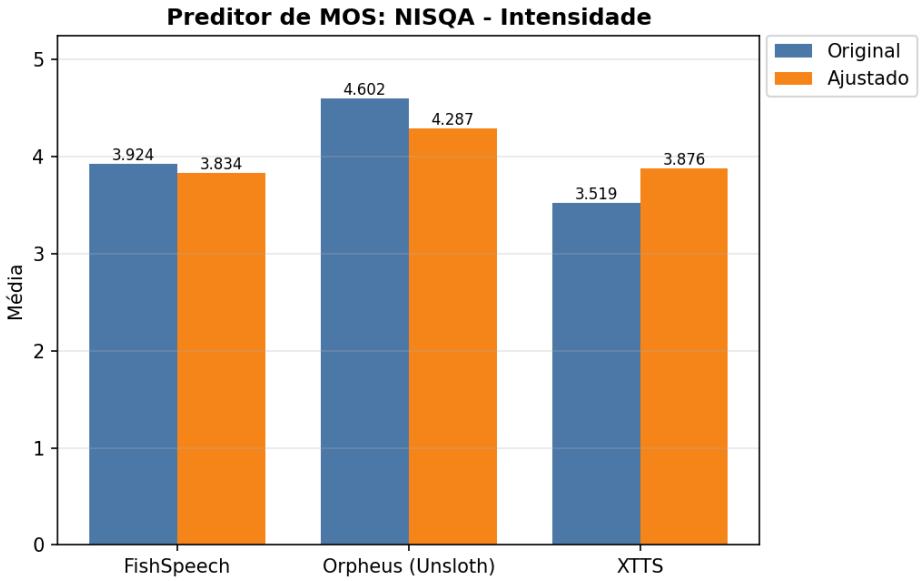
Figura C.24: *NISQA coloration*: histograma por modelo e condição.



Fonte: elaborado pelo autor (2026).

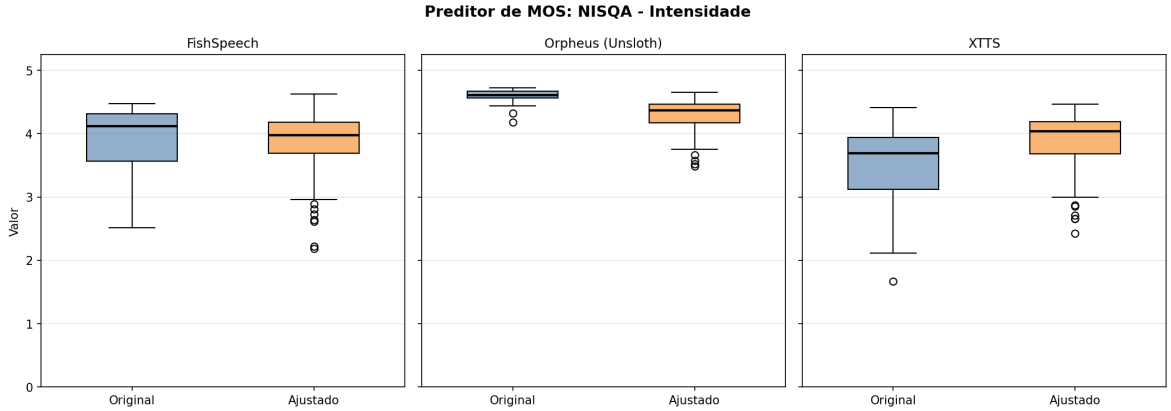
C.9 *NISQA*: Intensidade Sonora (*loudness*)

Figura C.25: *NISQA loudness*: médias por modelo e condição.



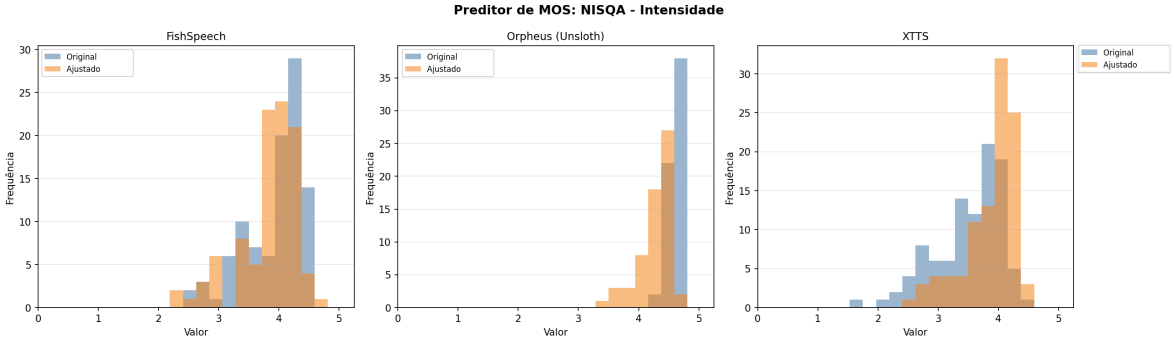
Fonte: elaborado pelo autor (2026).

Figura C.26: *NISQA loudness*: distribuição por modelo e condição.



Fonte: elaborado pelo autor (2026).

Figura C.27: *NISQA loudness*: histograma por modelo e condição.



Fonte: elaborado pelo autor (2026).

Apêndice D

Taxas de Erro de Caractere e de Palavra (CER e WER)

Este apêndice apresenta, de forma complementar à similaridade de Levenshtein utilizada no corpo da dissertação (Seção 5.4.2), as duas taxas de erro convencionalmente reportadas na literatura de reconhecimento automático de fala: a taxa de erro de caractere (*Character Error Rate*, CER) e a taxa de erro de palavra (*Word Error Rate*, WER) (Hunt, 1990; Morris et al., 2004).

As três métricas derivam da mesma distância de edição (Levenshtein, 1966) e incidem exatamente sobre os mesmos pares (texto enviado ao sintetizador, transcrição obtida pelo Whisper), já registrados na execução original da avaliação automática; não houve, portanto, nova transcrição nem nova síntese. A diferença está apenas na normalização aplicada à distância, conforme a Equação D.1:

$$\mathrm{sim}_{\mathrm{Lev}} = 1 - \frac{d_{\mathrm{car}}(\mathrm{ref}, \mathrm{hip})}{\max(|\mathrm{ref}|, |\mathrm{hip}|)}, \qquad \mathrm{CER} = \frac{d_{\mathrm{car}}(\mathrm{ref}, \mathrm{hip})}{|\mathrm{ref}|}, \qquad \mathrm{WER} = \frac{d_{\mathrm{pal}}(\mathrm{ref}, \mathrm{hip})}{N_{\mathrm{pal}}(\mathrm{ref})}, \tag{D.1}$$

em que d_{car} é a distância de edição em caracteres, d_{pal} a distância de edição em palavras, $|\cdot|$ o comprimento em caracteres e N_{pal} o número de palavras. A similaridade de Levenshtein é limitada ao intervalo $[0, 1]$ e cresce com o acerto; CER e WER são taxas de erro, não limitadas superiormente e decrescentes com o acerto. A normalização pelo comprimento da maior das duas cadeias, adotada na similaridade, torna a métrica simétrica e insensível a

transcrições muito mais longas que a referência, ao passo que CER e WER normalizam pela referência, o que é a convenção da área e permite comparação com valores publicados por outros trabalhos.

O texto foi normalizado da mesma forma nas três métricas (conversão para minúsculas, substituição de pontuação por espaço e colapso de espaços consecutivos), e os pares considerados são os mesmos reportados na Tabela 5.13. A Tabela D.1 apresenta os resultados, com o mesmo teste de postos sinalizados de Wilcoxon (Wilcoxon, 1945) pareado por índice de sentença empregado no corpo do texto.

Tabela D.1: Similaridade de Levenshtein, CER e WER por modelo: original vs. ajustado.

Modelo	**Métrica**	**Original**	**Ajustado**	Δ	p
XTTS	Similaridade de Levenshtein	0,999	0,904	$-0{,}095^{*}$	$< 0{,}001$
	CER	0,002	0,124	$+0{,}123^{*}$	$< 0{,}001$
	WER	0,004	0,236	$+0{,}232^{*}$	$< 0{,}001$
Fish Speech	Similaridade de Levenshtein	0,906	0,927	$+0{,}021$	0,455
	CER	0,101	0,075	$-0{,}026$	0,415
	WER	0,182	0,139	$-0{,}043$	0,233
Orpheus TTS	Similaridade de Levenshtein	0,855	0,921	$+0{,}066^{*}$	0,001
	CER	0,145	0,079	$-0{,}066^{*}$	0,001
	WER	0,187	0,171	$-0{,}016$	0,291

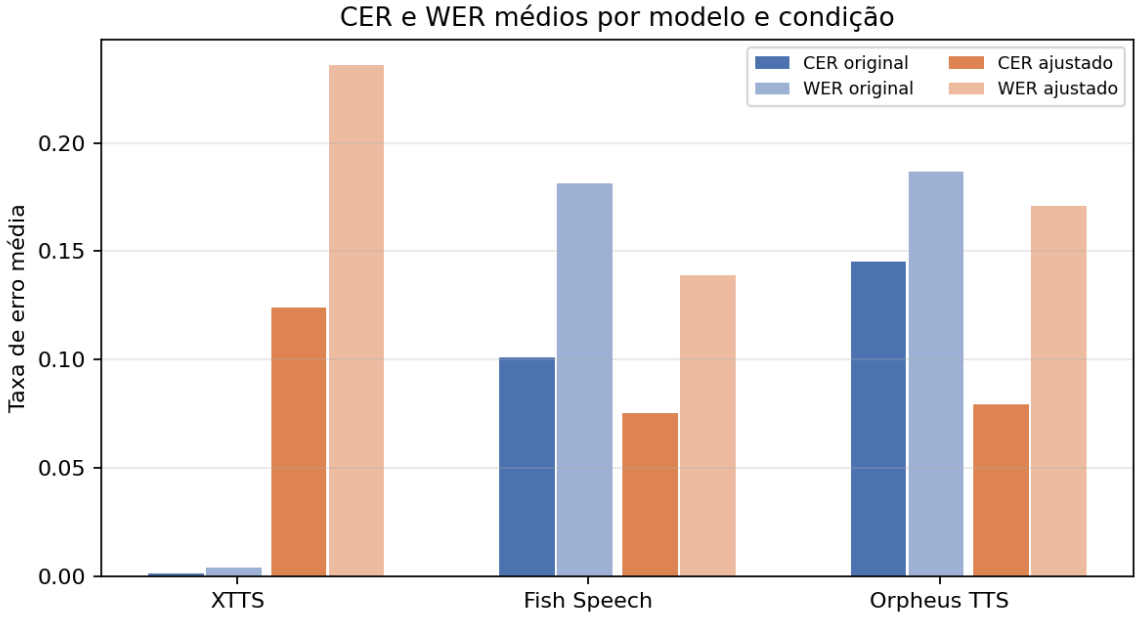
Para a similaridade de Levenshtein, $\Delta > 0$ indica vantagem do modelo ajustado; para CER e WER, por serem taxas de erro, $\Delta < 0$ indica vantagem do modelo ajustado. O asterisco marca diferenças significativas ($p < 0{,}05$).

Fonte: elaborada pelo autor (2026).

Os três indicadores são consistentes entre si quanto ao sentido do efeito e quanto à significância, o que era esperado, dado que derivam da mesma distância de edição. O modelo XTTS apresenta piora significativa nas três métricas, ainda que a partir de um patamar praticamente perfeito na condição original; já o modelo Fish Speech não apresenta diferença significativa em nenhuma delas; e o modelo Orpheus TTS apresenta melhora significativa em similaridade e em CER. A única divergência de leitura aparece no WER do modelo Orpheus TTS, cuja melhora não alcança significância estatística, ao contrário do que ocorre no CER do mesmo modelo: isso indica que os acertos adicionais do modelo ajustado concentram-se

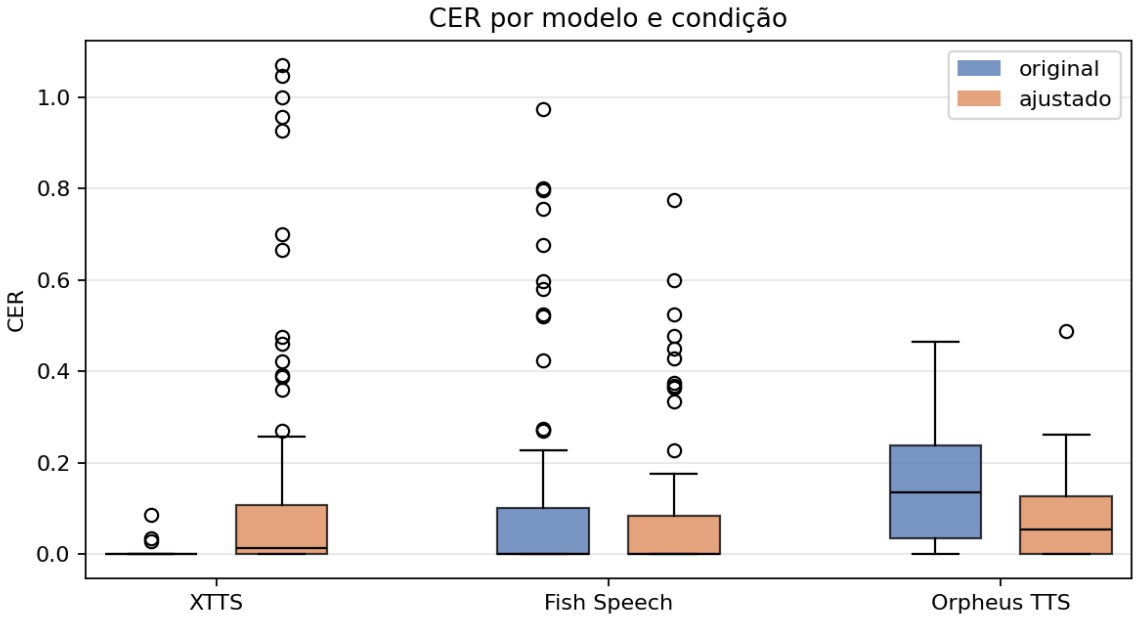
em caracteres dentro de palavras que permanecem erradas como um todo, situação em que o CER registra o ganho e o WER não. Vale registrar que os valores de CER e WER aqui reportados devem ser lidos com a ressalva discutida na Seção 5.6.3: eles medem o desempenho do reconhecedor sobre a fala sintetizada, e o reconhecedor empregado não tem exposição documentada à variedade regional em questão.

Figura D.1: CER e WER médios por modelo e condição (azul: original; laranja: ajustado).



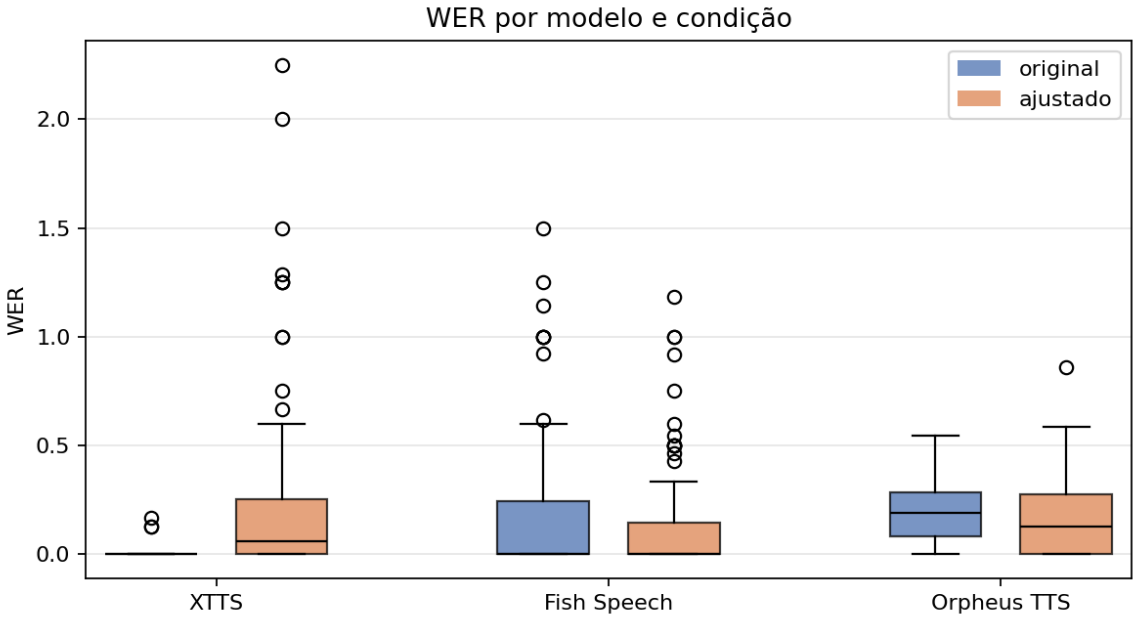
Fonte: elaborado pelo autor (2026).

Figura D.2: CER: distribuição por modelo e condição (azul: original; laranja: ajustado).



Fonte: elaborado pelo autor (2026).

Figura D.3: WER: distribuição por modelo e condição (azul: original; laranja: ajustado).



Fonte: elaborado pelo autor (2026).

Apêndice E

Parecer do Comitê de Ética em Pesquisa

Este apêndice reproduz, na íntegra, o parecer consubstanciado emitido pelo Comitê de Ética em Pesquisa (CEP), que aprovou a realização da pesquisa. O documento foi obtido por meio da Plataforma Brasil, plataforma nacional que centraliza a submissão e a tramitação de projetos de pesquisa envolvendo seres humanos, conforme referenciado na Seção 4.7.

UFCG - HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO DA UNIVERSIDADE FEDERAL DE CAMPINA GRANDE / HUAC - UFCG

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Incorporando Regionalismos e Sotaques em Modelos de Síntese de Fala
Pesquisador: THOMAZ DINIZ PINTO DE MORAIS
Área Temática:
Versão: 2
CAAE: 99217226.4.0000.5182
Instituição Proponente: UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 8.608.717

Apresentação do Projeto:
A presente pesquisa investiga a incorporação de regionalismos e sotaques em modelos de síntese de fala, com foco no sotaque paraibano. O estudo possui caráter experimental e aplicado, buscando analisar estratégias de adaptação de modelos de síntese de fala para representação de características regionais da fala brasileira.

Objetivo da Pesquisa:
Objetivo Primário:

Investigar estratégias de incorporação de regionalismos e sotaques em modelos de síntese de fala, com foco na adaptação de modelos para representação do sotaque paraibano, por meio da construção de uma base de dados regionalizada e da avaliação perceptual de áudios sintetizados.

Objetivo Secundário:

Construir uma base de dados de falas representativas do sotaque paraibano a partir da integração de bases públicas e coleta colaborativa de áudios.

Endereço: CAESE - Rua Dr. Chateaubriand, s/n.
Bairro: São José **CEP:** 58.107-670
UF: PB **Município:** CAMPINA GRANDE
Telefone: (83)2101-5545 **Fax:** (83)2101-5523 **E-mail:** cep@huac.ufcg.edu.br

UFCG - HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO DA UNIVERSIDADE FEDERAL DE CAMPINA GRANDE / HUAC - UFCG

Avaliação dos Riscos e Benefícios:

Os riscos são mínimos aos participantes, uma vez que envolve apenas o envio voluntário de gravações de fala e a participação em formulários online de avaliação perceptual de áudios sintetizados. Também pode-se pontuar certo desconforto relacionado ao uso da própria voz nas gravações enviadas durante a coleta colaborativa, bem como eventual cansaço leve durante a escuta e avaliação dos áudios presentes nos formulários. Existe também risco mínimo relacionado à privacidade dos dados coletados. Para minimizar esse risco, os áudios e respostas serão utilizados exclusivamente para fins acadêmicos, sem divulgação de identificação pessoal dos participantes. Os dados serão armazenados de forma restrita aos pesquisadores envolvidos no projeto.

Quanto aos benefícios, a pesquisa poderá contribuir para o desenvolvimento de tecnologias de síntese de fala mais representativas da diversidade linguística brasileira, especialmente no contexto de sotaques regionais da paraíba ainda pouco explorados em bases públicas e modelos computacionais. Espera-se também que o estudo contribua para a construção de recursos acadêmicos voltados ao sotaque paraibano, auxiliando futuras pesquisas nas áreas de síntese de fala, processamento de fala, inteligência artificial e linguística computacional.

Além dos benefícios científicos e tecnológicos, a pesquisa poderá auxiliar na valorização e representação de características linguísticas regionais em tecnologias de fala.

Comentários e Considerações sobre a Pesquisa:

Trata-se de uma pesquisa interessante, cujo resultado trará contribuições não apenas voltadas para questões tecnológicas; mas, em especial, para a alteridade, identidade e variação linguísticas, uma vez que se observará a visibilidade e a valorização do sotaque paraibano.

Considerações sobre os Termos de apresentação obrigatória:

Toda a documentação foi satisfatoriamente apresentada.

1-Informações Básicas do Projeto;
2-Termo de Compromisso dos Pesquisadores;
3-Instrumento de Coleta;
4-Termos de Consentimentos Livres e Esclarecidos - TCLEs;

Endereço: CAESE - Rua Dr. Chateaubriand, s/n.
Bairro: São José **CEP:** 58.107-670
UF: PB **Município:** CAMPINA GRANDE
Telefone: (83)2101-5545 **Fax:** (83)2101-5523 **E-mail:** cep@huac.ufcg.edu.br

UFCG - HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO DA UNIVERSIDADE FEDERAL DE CAMPINA GRANDE / HUAC - UFCG

5-Folha de Rosto;

6-Cronograma;

7-Orçamento;

8-Projeto.

Recomendações:

Não há recomendações.

Conclusões ou Pendências e Lista de Inadequações:

Protocolo de pesquisa apto para iniciar, após apreciação deste colegiado, considerando norma éticas vigentes.

Considerações Finais a critério do CEP:

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_2811331.pdf	15/07/2026 09:52:48		Aceito
Outros	termo_de_compromisso_assinado.pdf	19/06/2026 09:12:42	THOMAZ DINIZ PINTO DE MORAIS	Aceito
Outros	instrumento_coleta.pdf	19/06/2026 09:11:28	THOMAZ DINIZ PINTO DE MORAIS	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	tcle_coleta.pdf	19/06/2026 09:11:02	THOMAZ DINIZ PINTO DE MORAIS	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	tcle_avaliacao.pdf	19/06/2026 09:10:24	THOMAZ DINIZ PINTO DE MORAIS	Aceito
Folha de Rosto	folhaderosto.pdf	18/06/2026 19:02:03	THOMAZ DINIZ PINTO DE MORAIS	Aceito
Cronograma	cronograma.pdf	06/06/2026 08:57:02	THOMAZ DINIZ PINTO DE MORAIS	Aceito
Orçamento	orcamento.pdf	06/06/2026 08:55:56	THOMAZ DINIZ PINTO DE MORAIS	Aceito
Projeto Detalhado / Brochura	projeto_detalhado.pdf	06/06/2026 08:32:14	THOMAZ DINIZ PINTO DE MORAIS	Aceito

Endereço: CAESE - Rua Dr. Chateaubriand, s/n.
Bairro: São José **CEP:** 58.107-670
UF: PB **Município:** CAMPINA GRANDE
Telefone: (83)2101-5545 **Fax:** (83)2101-5523 **E-mail:** cep@huac.ufcg.edu.br

UFCG - HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO DA UNIVERSIDADE FEDERAL DE CAMPINA GRANDE / HUAC - UFCG

Continuação do Parecer: 8.608.717

Investigador	projeto_detalhado.pdf	06/06/2026 08:32:14	THOMAZ DINIZ PINTO DE MORAIS	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

CAMPINA GRANDE, 27 de Julho de 2026

Assinado por:
Andréia Oliveira Barros Sousa
(Coordenador(a))

Endereço: CAESE - Rua Dr. Chateaubriand, s/n.
Bairro: São José **CEP:** 58.107-670
UF: PB **Município:** CAMPINA GRANDE
Telefone: (83)2101-5545 **Fax:** (83)2101-5523 **E-mail:** cep@huac.ufcg.edu.br

Apêndice F

Instrumento de Coleta da Avaliação Perceptiva

Este apêndice reproduz, na íntegra, o formulário aplicado na avaliação perceptiva com ouvintes descrita na Seção 4.7. O instrumento é *on-line* e autoaplicável: inicia pelo Termo de Consentimento Livre e Esclarecido (TCLE) e pela pergunta de caracterização quanto à familiaridade com o sotaque paraibano, seguidos dos pares de áudios e das duas perguntas associadas a cada par, referentes à percepção de sotaque e à clareza e naturalidade. Em todos os itens, os áudios são apresentados de forma anonimizada, identificados apenas como "Áudio A" e "Áudio B", conforme descrito na Seção 4.7.5.

Avaliação de Sotaque Paraibano em Falas Sintetizadas

Esta é uma pesquisa acadêmica sobre percepção do sotaque paraibano em áudios sintetizados. Você ouvirá pares de áudios e indicará qual deles soa mais paraibano. A pesquisa é rápida e anônima.

* Indica uma pergunta obrigatória

1. **Consentimento para Participação** *

   O Termo de Consentimento Livre e Esclarecido (TCLE) desta pesquisa encontra-se disponível no link abaixo. Leia atentamente o documento antes de prosseguir.

   Clique neste link para obter acesso ao TCLE

   Marque todas que se aplicam.

   ☐ Li o TCLE e concordo voluntariamente em participar desta pesquisa.

Familiaridade com o sotaque paraibano

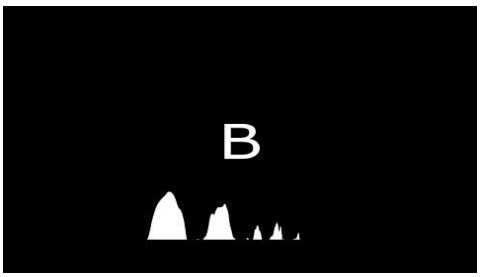
2. Você é da paraíba ou é familiarizado(a) com o sotaque paraibano? *

   Marcar apenas uma oval.

   ◯ Sou da paraíba
   ◯ Não, mas sou familiarizado(a) com o sotaque paraibano.
   ◯ Não, mas sou familiarizado(a) com sotaques nordestinos.
   ◯ Não, mas tenho pouca familiaridade.
   ◯ Não tenho familiaridade alguma com o sotaque

No vídeo a seguir, serão apresentados os áudios A e B. Ouça ambos atentamente antes de responder às perguntas.

Texto: Depois de muito procurar, finalmente encontrei as chaves que havia perdido durante a mudança.
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=1raj1CJ9COg

3. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

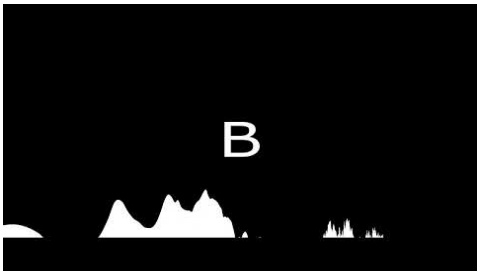
4. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Quando chegamos ao local do evento, a apresentação já havia começado e quase não encontramos lugares disponíveis.

Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=U4V8Kh6JgHQ

5. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

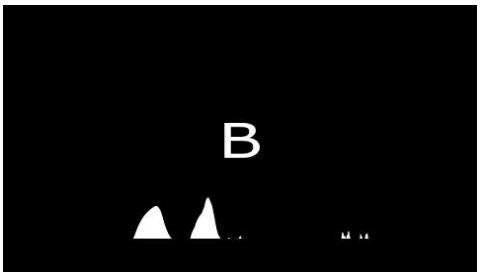
6. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Eu precisava apenas de alguns minutos para terminar aquela tarefa.

Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=KpJIM5Yfi3s

7. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

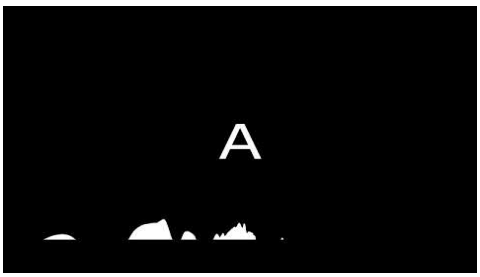
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

8. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Você pode repetir a última informação?
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=vdqTWwzpwbE

9. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

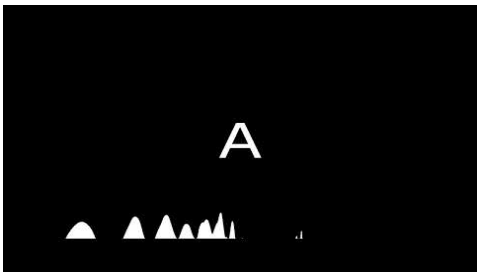
10. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Qual foi a primeira cidade que você visitou durante as férias?

Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=r2N6QWJJSLM

11. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

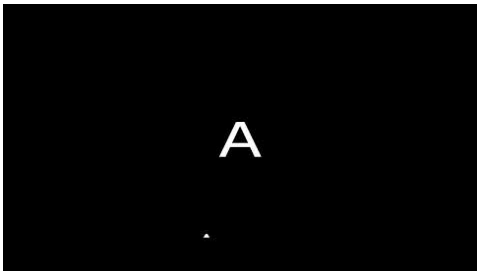
12. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Depois de meses de planejamento e preparação, a equipe finalmente iniciou a execução do projeto.

Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=UARzguMZQcQ

13. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

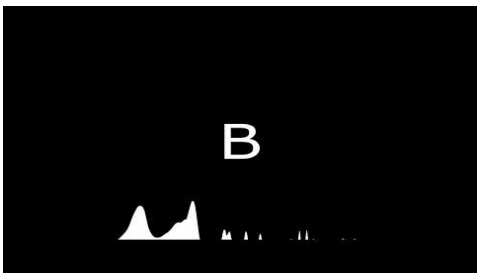
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

14. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Depois de comparar diferentes alternativas, escolhemos a solução que parecia mais adequada para o problema.
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=8Y-YZ5u6_Q0

15. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

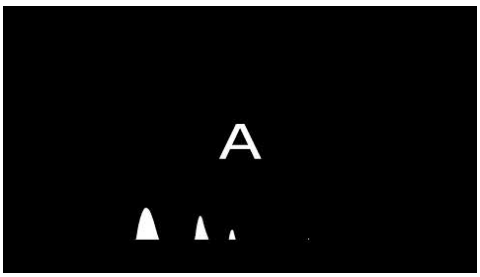
- Áudio A
- Áudio B
- Ambos aparentam ter sotaque paraibano
- Não consigo distinguir

16. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- Áudio A
- Áudio B
- Ambos soam igualmente bem
- Não consigo distinguir

Texto: Quando chegamos ao local do evento, a apresentação já havia começado e quase não encontramos lugares disponíveis.
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=urfCCocgbqU

17. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

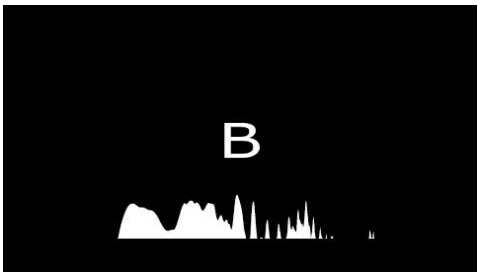
18. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: O funcionário organizou os documentos para facilitar futuras consultas.

Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=UsZvPaW6-D0

19. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

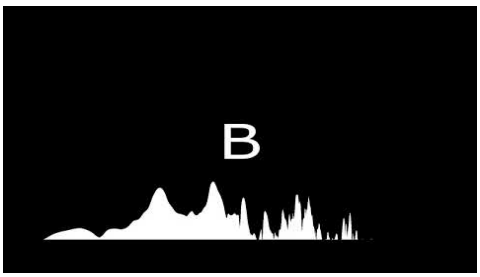
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

20. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: O café estava tão quente que precisei esperar alguns minutos antes de beber. Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=RoMEcBJo-Pk

21. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

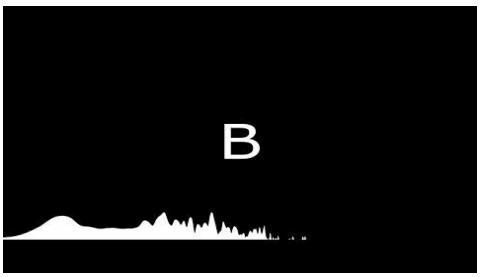
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

22. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: A praia estava praticamente vazia naquela manhã de terça-feira.
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=ow4OXKxtNy4

23. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

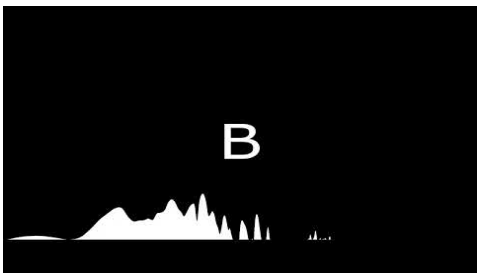
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

24. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Depois de meses de planejamento e preparação, a equipe finalmente iniciou a execução do projeto.
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=l8thtZ8TYDs

25. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

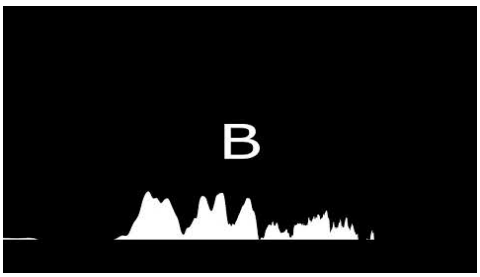
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

26. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Qual foi a primeira cidade que você visitou durante as férias?
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=eZesbOmkkbI

27. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

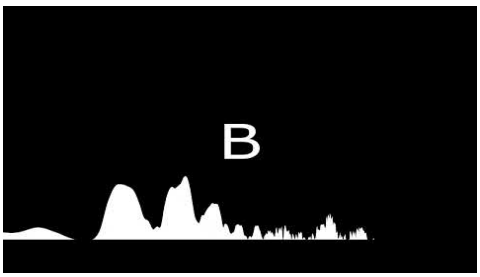
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

28. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Depois de vários testes, conseguimos identificar a origem do problema. Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=qhm_HsQ77O8

29. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

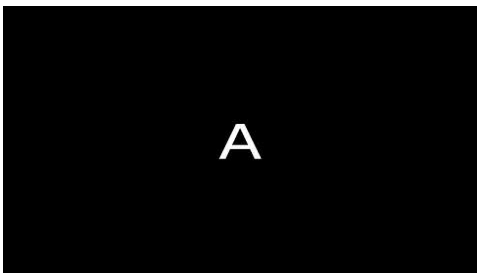
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

30. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: O relógio marcou exatamente sete horas da manhã.
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=S-Wp2vxZyuU

31. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

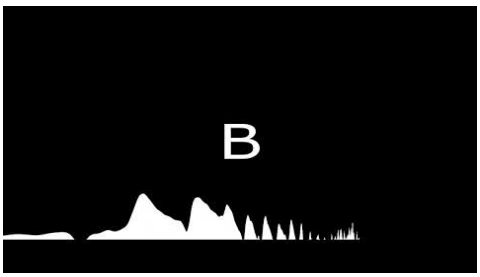
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

32. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Na semana passada houve uma queda de energia que afetou vários bairros da cidade.
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=loGxmlQh-us

33. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

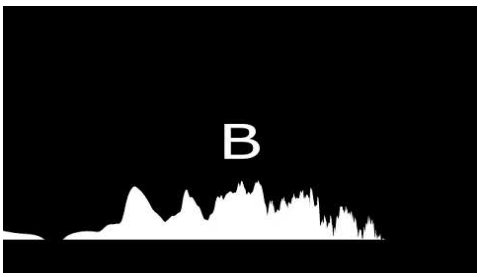
- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

34. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Texto: Depois de comparar diferentes alternativas, escolhemos a solução que parecia mais adequada para o problema.
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=0iyR0e3ZPKw

35. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

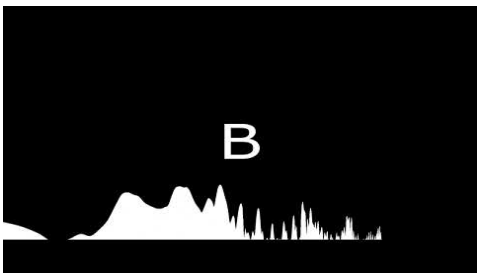
- Áudio A
- Áudio B
- Ambos aparentam ter sotaque paraibano
- Não consigo distinguir

36. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- Áudio A
- Áudio B
- Ambos soam igualmente bem
- Não consigo distinguir

Texto: O visitante demonstrou bastante interesse pelas atividades desenvolvidas no laboratório.
Ouça os áudios e responda às perguntas abaixo.



http://youtube.com/watch?v=Ss3fZOFrKbI

37. Qual áudio você considera que representa melhor o sotaque paraibano? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos aparentam ter sotaque paraibano
- ◯ Não consigo distinguir

38. Considerando a naturalidade e a clareza da fala, qual áudio você considera melhor? *

Marcar apenas uma oval.

- ◯ Áudio A
- ◯ Áudio B
- ◯ Ambos soam igualmente bem
- ◯ Não consigo distinguir

Este conteúdo não foi criado nem aprovado pelo Google.