

GILSON MACIEL

RECONHECIMENTO DE FALA, TRANSCRIÇÃO E INTELIGÊNCIA ARTIFICIAL

Teoria, esquemas e questões comentadas

400 questões comentadas
conforme o edital de 2026

1ª EDIÇÃO · 2026



Gilson Maciel

RECONHECIMENTO DE FALA, TRANSCRIÇÃO E INTELIGÊNCIA ARTIFICIAL

Teoria, esquemas e questões comentadas

2026

APRESENTAÇÃO

.....

O registro dos debates é uma das funções mais antigas e mais sensíveis de um parlamento. É por meio dele que a palavra dita em plenário e nas comissões se torna documento público, consultável, citável e preservado, base da transparência, da memória institucional e do controle social. Durante mais de um século, essa tarefa repousou sobre a taquigrafia e a revisão humana. Nos últimos anos, o reconhecimento automático de fala e os grandes modelos de linguagem transformaram o trabalho: o áudio passou a ser convertido em texto por máquinas, e o profissional do registro passou a revisar, corrigir, formatar e validar o que a máquina produz, sem abrir mão da fidelidade ao que foi dito.

O edital do concurso da Câmara dos Deputados de 2026 para o cargo de Analista Legislativo, na especialidade Registro e Redação, reconheceu essa mudança ao incluir o item *Reconhecimento de fala, transcrição e inteligência artificial*. O conteúdo é amplo e heterogêneo: vai da acústica e da arquitetura dos modelos à prática da degravação, das normas de formatação das notas taquigráficas à gestão da qualidade, da recuperação de informação à proteção de dados pessoais, dos direitos autorais à segurança da informação. São raras as obras que reúnem esses temas na medida e na linguagem que a prova exige, e este livro procura preencher essa lacuna.

A organização da obra

O livro divide-se em quatro partes e doze capítulos. A **Parte I** trata da tecnologia do reconhecimento de fala: os fundamentos (sinal, atributos acústicos, modelos acústicos e modelos de língua, métricas como a taxa de erro de palavras, a WER), os modelos de ponta a ponta e os grandes modelos de linguagem (LLMs) que recebem áudio, a diarização, a pontuação e o processamento em tempo real, e as particularidades do português brasileiro, com as principais ferramentas e serviços disponíveis.

A **Parte II** cuida da transcrição como atividade profissional: de gravação e revisão humana, retextualização, formatação e normalização das notas taquigráficas, ferramentas de apoio, controle de qualidade, níveis de serviço e adaptação de vocabulário.

A **Parte III** examina os usos da inteligência artificial no trabalho parlamentar: indexação, busca e geração aumentada por recuperação, elaboração de atas e resumos e modelos de linguagem em português, sempre à luz da experiência da própria Câmara dos Deputados e das diretrizes da União Interparlamentar. A **Parte IV** reúne o marco jurídico e de governança: a Lei Geral de Proteção de Dados Pessoais, com o Ato da Mesa nº 152/2020, que a regulamenta na Câmara; os direitos autorais, a ética e a governança da IA, com destaque para o Ato da Mesa nº 225/2025 e a Portaria nº 227/2025; e a segurança da informação em cadeias de processamento (*pipelines*) de áudio e transcrição.

A estrutura de cada capítulo

Todos os capítulos seguem o mesmo roteiro. Abrem com uma apresentação do tema, os objetivos de aprendizagem, um quadro de termos e siglas e um esquema das seções. O texto é entremeado de quadros comparativos, de boxes de conceito e de dicas, de palavras-chave para revisão e, sobretudo, de boxes intitulados *Onde o item pode errar*, que mostram como as bancas costumam transformar uma afirmação correta em incorreta: troca de conceitos, inversão de números, generalizações indevidas. Cada capítulo termina com uma síntese e com questões comentadas.

O primeiro bloco de questões reúne itens de concursos anteriores, selecionados entre os que efetivamente cobram o conteúdo do capítulo, com prioridade para os da banca Cebraspe e para as provas recentes, inclusive as da própria Câmara dos Deputados em 2026. O segundo bloco traz questões inéditas, elaboradas no formato de julgamento de itens (certo ou errado) próprio dessa banca e graduadas em quatro níveis de dificuldade. Ao todo, a obra contém 400 questões comentadas, das quais 276 inéditas e 124 de concursos anteriores. Todos os comentários indicam a seção do capítulo em que o assunto é tratado.

O livro foi escrito para ser autossuficiente. Os conceitos de aprendizado de máquina, redes neurais e grandes modelos de linguagem,

que sustentam boa parte do conteúdo, são apresentados desde a base nos Capítulos 1 e 2; os termos técnicos são explicados no ponto em que aparecem ou, quando tratados em outro capítulo, indicados por remissão; e, ao final, uma **lista de siglas** desdobra todas as siglas usadas na obra e um **glossário** reúne cerca de duzentas definições para consulta rápida.

Correspondência entre o edital e os capítulos

O quadro a seguir indica onde cada subitem do edital nº 1 – CÂMARA DOS DEPUTADOS – ANALISTA LEGISLATIVO, DE 2 DE OUTUBRO DE 2026 é tratado. Os itens do edital são reproduzidos como no original, inclusive os termos em inglês: *end-to-end* (de ponta a ponta), *human in the loop* (humano no circuito), *language model adaptation* (adaptação do modelo de língua) e *pipelines* (cadeias de processamento). As siglas estão desdobradas na lista de siglas, ao final do livro. Alguns temas reaparecem em mais de um capítulo, como a proteção de dados nas ferramentas em nuvem e a revisão humana nas aplicações de IA generativa (a que produz conteúdo novo, como textos e resumos; Capítulo 1, seção 4.4).

■ QUADRO A
EDITAL E CAPÍTULOS

Item do edital	Capítulo e seção
1.1 Reconhecimento automático de fala (ASR)	Capítulo 1
1.2 Modelos de fala <i>end-to-end</i>	Capítulo 2, seções 2 a 5
1.3 LLMs multimodais de áudio nativo	Capítulo 2, seção 6
1.4 Diarização de falantes	Capítulo 3, seção 2
1.5 Pontuação automática e capitalização	Capítulo 3, seção 3
1.6 Desafios do ASR em português brasileiro	Capítulo 4, seções 2 e 3
1.7 Transcrição em tempo real	Capítulo 3, seção 4
1.8 Ferramentas e APIs (Whisper, Google, Azure, AWS)	Capítulo 4, seções 4 e 5
2.1 Transcrição automática, gravação e revisão humana	Capítulo 5, seção 2
2.2 Gravação assistida por IA (<i>human in the loop</i>)	Capítulo 5, seção 3
2.3 Revisão: critérios de qualidade e fidelidade	Capítulo 5, seção 4
2.4 Formatação de transcrições parlamentares	Capítulo 6, seção 2
2.5 Normalização de texto	Capítulo 6, seções 3 a 5

Item do edital	Capítulo e seção
2.6 Ferramentas de CAT e editores	Capítulo 7, seção 2
2.7 Controle de qualidade, auditoria de amostras, SLA	Capítulo 7, seções 3 e 4
2.8 Vocabulário técnico e <i>language model adaptation</i>	Capítulo 7, seção 5
3.1 Aplicações de IA em parlamentos	Capítulo 8, seções 2 e 3
3.2 Busca semântica em acervos parlamentares	Capítulo 8, seções 4 a 6
3.3 Geração automática de atas e resumos	Capítulo 9, seções 2 a 4
3.4 Modelos multilíngues e especializados em português	Capítulo 9, seção 5
4.1 LGPD e proteção de dados em transcrição e IA	Capítulo 10
4.2 Direitos autorais e ética no uso de IA	Capítulo 11
4.3 Segurança da informação em <i>pipelines</i> de áudio	Capítulo 12

Critérios adotados

Três critérios orientaram a redação. O primeiro foi a **fidelidade às fontes primárias**: leis, atos normativos, normas técnicas e artigos científicos foram consultados diretamente, e as informações sobre a experiência da Câmara dos Deputados foram extraídas de documentos institucionais. O segundo foi a **atualidade**: o texto reflete a legislação e o estado da técnica até outubro de 2026, inclusive a transformação da Autoridade Nacional de Proteção de Dados (ANPD) em Agência Nacional de Proteção de Dados, pela Medida Provisória nº 1.317/2025, convertida na Lei nº 15.352/2026, e a tramitação do Projeto de Lei nº 2.338/2023. Como recursos de produtos comerciais mudam com frequência, o livro privilegia os conceitos estáveis e indica, quando necessário, que determinada funcionalidade se refere a uma versão ou a uma data. O terceiro foi a **adequação à prova**: a profundidade técnica foi calibrada para o que um analista do registro precisa compreender e o que uma banca costuma cobrar, sem exigir formação prévia em computação.

Como utilizar este livro

O livro foi pensado para dois tipos de leitor. Quem chega ao tema sem conhecimentos prévios sugere-se ler os capítulos na ordem, pois os primeiros fornecem o vocabulário dos seguintes: os boxes de conceito explicam as noções de base (vetor, modelo generativo, interface de

programação, conceitos jurídicos e de segurança) no ponto em que elas se tornam necessárias, e as remissões indicam onde cada termo foi explicado. Quem já domina parte da matéria pode usar o livro para revisão ou consulta pontual, começando pelo quadro de correspondência com o edital e conferindo nas questões as distinções que as bancas exploram. Na primeira leitura, convém resolver as questões ao final de cada capítulo antes de consultar os comentários; na revisão, os quadros de palavras-chave, as sínteses e os boxes *Onde o item pode errar* permitem percorrer rapidamente todo o conteúdo. A lista de siglas e o glossário servem à consulta rápida dos termos técnicos, e as referências indicam as obras para aprofundamento. Que o material seja útil a quem se prepara para integrar uma equipe de registro de longa tradição, num momento em que a tecnologia redefine, mas não substitui, o papel do profissional que garante a fidelidade da palavra parlamentar.

O Autor

SUMÁRIO

AGRADECIMENTOS ERRO! INDICADOR NÃO DEFINIDO.

APRESENTAÇÃO 7

RECONHECIMENTO DE FALA E TRANSCRIÇÃO DE ÁUDIO 20

CAPÍTULO 1 FUNDAMENTOS DO RECONHECIMENTO AUTOMÁTICO DE FALA

..... 21

1. INTRODUÇÃO 23

1.1. Como ler os itens desta disciplina 24

2. O QUE É RECONHECIMENTO AUTOMÁTICO DE FALA 24

2.1. Definição e tarefas vizinhas 24

2.2. Breve histórico 25

3. A FALA COMO SINAL 27

3.1. Do som ao sinal digital 27

3.2. Noções de fonética úteis ao ASR 29

3.3. Atributos acústicos: do sinal ao espectrograma 29

4. A ARQUITETURA CLÁSSICA DO RECONHECIMENTO 33

4.1. A equação fundamental 33

4.2. Os componentes 34

4.3. Por que a arquitetura clássica ainda importa 37

4.4. Noções de aprendizado de máquina e de redes neurais 38

5. COMO SE MEDE A QUALIDADE DO RECONHECIMENTO 43

5.1. A taxa de erro de palavras (WER) 43

5.2. Outras métricas 45

5.3. A WER no mundo real: o caso da Câmara 46

6. FATORES QUE AFETAM O DESEMPENHO 48

6.1. Fatores acústicos 48

6.2. Fatores linguísticos e do falante 49

6.3. Como os sistemas ganham robustez 49

SÍNTESE DO CAPÍTULO 51

QUESTÕES DE CONCURSO COMENTADAS 52

QUESTÕES INÉDITAS COMENTADAS 53

CAPÍTULO 2 MODELOS END-TO-END E LLMS DE ÁUDIO NATIVO 63

1. INTRODUÇÃO 65

2. O QUE É UM MODELO DE PONTA A PONTA 65

2.1. Uma rede única, otimizada de uma vez 65

2.2. Vantagens e custos 66

3. AS TRÊS FAMÍLIAS DE MODELOS DE PONTA A PONTA 67

3.1. CTC (<i>Connectionist Temporal Classification</i>)	68
3.2. RNN-T (transdutor)	68
3.3. <i>Encoder-decoder</i> com atenção (AED)	69
4. ENCODERS E APRENDIZADO AUTO-SUPERVISIONADO	71
4.1. Do recorrente ao Conformer	71
4.2. Aprender com áudio sem transcrição	73
5. O WHISPER	75
5.1. Supervisão fraca em larga escala	75
5.2. Arquitetura e formato multitarefa	75
5.3. Versões e licença	76
5.4. Limitações conhecidas	77
6. LLMS MULTIMODAIS DE ÁUDIO NATIVO	80
6.1. O que é um grande modelo de linguagem	80
6.2. Cascata e modelo nativo	82
6.3. O que muda para a transcrição	83
SÍNTESE DO CAPÍTULO	85
QUESTÕES INÉDITAS COMENTADAS	86

CAPÍTULO 3 DIARIZAÇÃO, PONTUAÇÃO E TRANSCRIÇÃO EM TEMPO REAL

.....	95
1. INTRODUÇÃO	96
2. DIARIZAÇÃO DE FALANTES	97
2.1. O que é e o que não é	97
2.2. O <i>pipeline</i> modular	98
2.3. Diarização neural de ponta a ponta	99
2.4. Como se mede: a DER	100
2.5. Desafios no ambiente parlamentar	101
3. PONTUAÇÃO E CAPITALIZAÇÃO AUTOMÁTICAS	102
3.1. O problema	102
3.2. Pistas textuais e prosódicas	103
3.3. Avaliação e limites	103
4. TRANSCRIÇÃO EM TEMPO REAL	105
4.1. Lote e fluxo contínuo	105
4.2. Métricas próprias	105
4.3. Modelos e adaptações	106
4.4. Legendagem ao vivo e acessibilidade	107
SÍNTESE DO CAPÍTULO	108
QUESTÕES DE CONCURSO COMENTADAS	109
QUESTÕES INÉDITAS COMENTADAS	110

CAPÍTULO 4 O ASR EM PORTUGUÊS BRASILEIRO E AS FERRAMENTAS DO

MERCADO	117
1. INTRODUÇÃO	119
2. DESAFIOS DO PORTUGUÊS BRASILEIRO	119
2.1. Recursos e <i>corpora</i>	119
2.2. Variação regional e social	120
2.3. Fala espontânea, homofonias e sândi	121
3. VOCABULÁRIO TÉCNICO E PARLAMENTAR	123

3.1. O que torna o discurso parlamentar difícil	123
3.2. O ambiente acústico do Parlamento	124
3.3. O que a experiência da Câmara mostra	124
4. FERRAMENTAS E APIS	125
4.1. Visão geral.....	125
4.2. Whisper (OpenAI).....	126
4.3. Google Cloud Speech-to-Text.....	127
4.4. Azure AI Speech (Microsoft)	128
4.5. Amazon Transcribe (AWS)	129
5. COMO ESCOLHER UMA SOLUÇÃO	131
5.1. Critérios	131
5.2. O caso da Câmara dos Deputados	132
SÍNTESE DO CAPÍTULO	133
QUESTÕES DE CONCURSO COMENTADAS	134
QUESTÕES INÉDITAS COMENTADAS	136

DEGRAVAÇÃO E PÓS-PROCESSAMENTO DE TRANSCRIÇÕES.....143

CAPÍTULO 5 TRANSCRIÇÃO, DEGRAVAÇÃO E REVISÃO HUMANA	145
1. INTRODUÇÃO	146
2. TRANSCRIÇÃO, DEGRAVAÇÃO, REVISÃO E REDAÇÃO FINAL	147
2.1. Os conceitos.....	147
2.2. Tipos de transcrição.....	148
2.3. Transcrição e retextualização em Marcuschi.....	148
3. DEGRAVAÇÃO ASSISTIDA POR IA COM HUMANO NO CIRCUITO	151
3.1. O conceito de humano no circuito	151
3.2. As etapas do fluxo	152
3.3. O viés de automação	153
4. CRITÉRIOS DE QUALIDADE E FIDELIDADE	155
4.1. Os critérios	155
4.2. Trechos incompreensíveis e dúvidas	156
4.3. Classificação dos erros	156
SÍNTESE DO CAPÍTULO	158
QUESTÕES INÉDITAS COMENTADAS	159

CAPÍTULO 6 FORMATAÇÃO PARLAMENTAR E NORMALIZAÇÃO DE TEXTO

.....	167
1. INTRODUÇÃO	168
2. A FORMATAÇÃO DAS NOTAS TAQUIGRÁFICAS	169
2.1. O cabeçalho.....	169
2.2. A identificação dos oradores	170
2.3. Apartes, interrupções e intervenções	170
2.4. Marcações de eventos	171
2.5. Outras convenções de transcrição	172
3. NORMALIZAÇÃO DE TEXTO	174
3.1. Conceito: TN e ITN	174
3.2. Categorias e ambiguidades.....	175
3.3. Abordagens	175

3.4. Normalizar para publicar e normalizar para medir	176
4. REGRAS USUAIS DE GRAFIA	177
4.1. Por que um manual.....	177
4.2. Números, datas, horas e valores.....	177
4.3. Siglas e abreviações.....	178
4.4. Referências normativas e dispositivos	178
4.5. Termos estrangeiros e técnicos	178
5. A NORMALIZAÇÃO NO FLUXO DE PÓS-PROCESSAMENTO	180
SÍNTESE DO CAPÍTULO	181
QUESTÕES INÉDITAS COMENTADAS	182

CAPÍTULO 7 FERRAMENTAS, CONTROLE DE QUALIDADE E VOCABULÁRIO

ADAPTADO	187
1. INTRODUÇÃO	189
2. FERRAMENTAS DE CAT E EDITORES DE TRANSCRIÇÃO	189
2.1. O sentido original de CAT	189
2.2. O editor de transcrição.....	190
3. CONTROLE DE QUALIDADE.....	192
3.1. Garantia e controle	192
3.2. Revisão em dupla	192
3.3. Auditoria por amostragem	192
3.4. Indicadores	194
4. ACORDOS DE NÍVEL DE SERVIÇO	195
4.1. O acordo de nível de serviço e a ITIL.....	195
4.2. Nível de serviço, utilidade, garantia, sustentabilidade e experiência	195
4.3. OLA, contrato de apoio e tipos de SLA	197
4.4. O que um SLA deve conter	198
4.5. Disponibilidade e indicadores de confiabilidade	199
4.6. SLA nos contratos públicos.....	199
5. VOCABULÁRIO TÉCNICO PARLAMENTAR E ADAPTAÇÃO DO RECONHECIMENTO	202
5.1. O vocabulário a adaptar.....	202
5.2. A adaptação clássica do modelo de língua	202
5.3. A adaptação nos modelos modernos	203
5.4. Manutenção contínua	205
SÍNTESE DO CAPÍTULO	206
QUESTÕES DE CONCURSO COMENTADAS	207
QUESTÕES INÉDITAS COMENTADAS	215

IA APLICADA NO CONTEXTO PARLAMENTAR..... 223

CAPÍTULO 8 IA EM PARLAMENTOS, INDEXAÇÃO E BUSCA SEMÂNTICA.....	225
1. INTRODUÇÃO	227
2. A INTELIGÊNCIA ARTIFICIAL NOS PARLAMENTOS	227
2.1. Aplicações	227
2.2. As Diretrizes da UIP para IA em parlamentos	228
2.3. O programa Ulysses, da Câmara dos Deputados	228
3. INDEXAÇÃO	229

3.1. Conceito	229
3.2. Vocabulários controlados	230
3.3. Extração automática	230
4. RECUPERAÇÃO DE INFORMAÇÃO CLÁSSICA	231
4.1. Índice invertido e modelos de relevância.....	231
4.2. Avaliação	233
4.3. Os limites da busca por palavras	233
5. BUSCA SEMÂNTICA.....	234
5.1. <i>Embeddings</i> e similaridade	234
5.2. Bancos vetoriais e busca aproximada	235
5.3. Busca híbrida e reclassificação	236
6. GERAÇÃO AUMENTADA POR RECUPERAÇÃO (RAG)	237
6.1. Conceito	237
6.2. As etapas.....	238
6.3. Vantagens e riscos	238
6.4. RAG no acervo de discursos.....	239
SÍNTESE DO CAPÍTULO	240
QUESTÕES INÉDITAS COMENTADAS	241

CAPÍTULO 9 ATAS, RESUMOS E MODELOS DE LINGUAGEM EM PORTUGUÊS

.....	249
1. INTRODUÇÃO	251
2. OS DOCUMENTOS	251
3. SUMARIZAÇÃO AUTOMÁTICA	252
3.1. Extrativa e abstrativa	252
3.2. Tipos de resumo	253
3.3. Avaliação	253
3.4. A particularidade das reuniões	254
4. GERAÇÃO AUTOMÁTICA DE ATAS E RESUMOS	255
4.1. Um fluxo possível.....	255
4.2. Riscos próprios	256
4.3. Salvaguardas	256
5. MODELOS MULTILÍNGUES E ESPECIALIZADOS EM PORTUGUÊS	257
5.1. Por que a língua importa.....	257
5.2. Modelos codificadores	258
5.3. Modelos geradores	258
5.4. Critérios de escolha	260
SÍNTESE DO CAPÍTULO	261
QUESTÕES INÉDITAS COMENTADAS	262

PROTEÇÃO E SEGURANÇA DA INFORMAÇÃO 269

CAPÍTULO 10 LGPD E PROTEÇÃO DE DADOS EM SISTEMAS DE

TRANSCRIÇÃO E IA.....	271
1. INTRODUÇÃO	274
2. FUNDAMENTOS, OBJETIVO E ÂMBITO DE APLICAÇÃO	275
2.1. Objetivo e fundamentos	275
2.2. Âmbito de aplicação.....	276

2.3. Exceções	276
3. CONCEITOS	278
3.1. Os conceitos do art. 5º	278
3.2. Anonimização e pseudonimização	279
3.3. A voz como dado pessoal e como dado biométrico.....	279
4. PRINCÍPIOS.....	280
5. BASES LEGAIS	282
5.1. Dados pessoais comuns (art. 7º)	282
5.2. Dados sensíveis (art. 11)	284
5.3. Orientações da ANPD para o Poder Público	285
5.4. Crianças e adolescentes (art. 14).....	286
5.5. Término do tratamento e conservação (arts. 15 e 16)	286
6. TRATAMENTO DE DADOS PELO PODER PÚBLICO	288
6.1. Regras gerais (arts. 23 a 32).....	288
6.2. Publicidade dos debates e proteção de dados	289
6.3. A LGPD na Câmara dos Deputados: o Ato da Mesa nº 152/2020	290
7. DIREITOS DO TITULAR E DECISÕES AUTOMATIZADAS	293
7.1. Os direitos do art. 18.....	293
7.2. Decisões automatizadas (art. 20).....	294
8. AGENTES, ENCARREGADO, RIPD E RESPONSABILIDADE	295
8.1. Controlador, operador e encarregado	295
8.2. Responsabilidade e ressarcimento (arts. 42 a 45).....	296
8.3. Transferência internacional de dados (arts. 33 a 36).....	298
9. SEGURANÇA, INCIDENTES, BOAS PRÁTICAS, SANÇÕES E ANPD	299
9.1. Segurança e incidentes.....	299
9.2. Boas práticas e governança.....	300
9.3. Sanções administrativas (art. 52).....	300
9.4. A ANPD	301
SÍNTESE DO CAPÍTULO	302
QUESTÕES DE CONCURSO COMENTADAS	304
QUESTÕES INÉDITAS COMENTADAS	341
CAPÍTULO 11 DIREITOS AUTORAIS, ÉTICA E GOVERNANÇA DA IA	349
1. INTRODUÇÃO	351
2. DIREITOS AUTORAIS	352
2.1. O que é protegido e o que não é.....	352
2.2. Autoria e direitos	352
2.3. Limitações e o caso dos discursos	353
3. IA E DIREITOS AUTORAIS	354
3.1. Quem é autor do conteúdo gerado por IA	354
3.2. Obras protegidas no treinamento de modelos.....	355
3.3. A Portaria nº 227/2025 e a propriedade intelectual	355
4. ÉTICA DA INTELIGÊNCIA ARTIFICIAL	356
4.1. Referências internacionais.....	356
4.2. Os temas centrais.....	356
5. O ATO DA MESA Nº 225/2025.....	357
5.1. Natureza e estrutura	357
5.2. Conteúdo	358
6. A PORTARIA Nº 227/2025	360

6.1. Natureza e definições	360
6.2. Princípios, objetivos e riscos	361
6.3. Diretrizes	361
6.4. Gestão de riscos e o CETIA.....	363
6.5. Papéis	365
7. O MARCO REGULATÓRIO DA IA	366
7.1. O PL nº 2.338/2023	366
7.2. O Regulamento europeu de IA.....	367
SÍNTESE DO CAPÍTULO	368
QUESTÕES INÉDITAS COMENTADAS	369

CAPÍTULO 12 SEGURANÇA DA INFORMAÇÃO EM *PIPELINES* DE ÁUDIO E

TRANSCRIÇÃO	377
1. INTRODUÇÃO	379
2. CONCEITOS	381
2.1. Propriedades da segurança da informação	381
2.2. Ameaça, vulnerabilidade, risco e incidente	381
2.3. A governança de segurança na Câmara	382
3. O <i>PIPELINE</i> E SUA SUPERFÍCIE DE ATAQUE.....	383
4. AMEAÇAS ESPECÍFICAS DE SISTEMAS DE FALA E DE IA.....	384
4.1. Falsificação de voz (<i>deepfakes</i>).....	384
4.2. Ataques adversariais ao reconhecimento de fala	385
4.3. Envenenamento e cadeia de suprimentos	385
4.4. Injeção de <i>prompt</i> e outras ameaças aos LLMs	385
4.5. Vazamento por provedores e uso não autorizado de ferramentas.....	386
5. CONTROLES	387
5.1. Classificação e minimização	387
5.2. Controle de acesso	388
5.3. Criptografia e integridade.....	388
5.4. Registros e trilhas de auditoria	389
5.5. Continuidade e cópias de segurança	389
5.6. Segurança em nuvem e contratos.....	389
5.7. Gestão de incidentes	390
6. NORMAS E REFERENCIAIS.....	391
6.1. ISO/IEC 27001 e 27002	391
6.2. Referenciais para IA	392
SÍNTESE DO CAPÍTULO	393
QUESTÕES DE CONCURSO COMENTADAS	394
QUESTÕES INÉDITAS COMENTADAS	400

SIGLAS.....	407
-------------	-----

GLOSSÁRIO	415
-----------------	-----

REFERÊNCIAS.....	429
------------------	-----

P A R T E I

RECONHECIMENTO DE FALA E TRANSCRIÇÃO DE ÁUDIO

.....

Capítulo 1 Fundamentos do reconhecimento automático de fala

Capítulo 2 Modelos end-to-end e LLMs de áudio

Capítulo 3 Diarização, pontuação e transcrição em tempo real

Capítulo 4 O ASR em português brasileiro e as ferramentas do mercado

CAPÍTULO 1

FUNDAMENTOS DO RECONHECIMENTO AUTOMÁTICO DE FALA

.....

Este capítulo apresenta o reconhecimento automático de fala (ASR) desde a base: como a voz se torna um sinal digital, como esse sinal é descrito por atributos acústicos, como a arquitetura clássica combina modelo acústico, léxico e modelo de língua e como se mede a qualidade de uma transcrição automática. Esses conceitos sustentam todos os capítulos seguintes e são o terreno preferido de itens que trocam a função de um componente pela de outro.

OBJETIVOS DO CAPÍTULO

- Explicar o que é reconhecimento automático de fala e distingui-lo de tarefas vizinhas, como identificação de locutor, compreensão de linguagem e síntese de fala.
- Descrever a conversão da voz em sinal digital (amostragem, quantização, taxa de amostragem) e a extração de atributos (janelas, espectrograma, escala mel, MFCC).
- Reconhecer a função do modelo acústico, do léxico, do modelo de língua e do decodificador na arquitetura clássica, e a equação que os combina.
- Compreender as noções básicas de aprendizado de máquina e de redes neurais usadas em todo o livro (treinamento, generalização, formas de aprendizado, tipos de rede, pré-treinamento e ajuste fino).
- Calcular a taxa de erro de palavras (WER) e interpretar seus limites, além de conhecer CER, SER, confiança, latência e fator de tempo real.
- Identificar os fatores acústicos, linguísticos e do falante que degradam o desempenho de um sistema de ASR.

■ QUADRO 1.1

TERMOS E SIGLAS DO CAPÍTULO

Termo ou sigla	Significado neste capítulo
ASR	<i>Automatic Speech Recognition</i> : reconhecimento automático de fala; conversão de fala em texto (também chamada STT, <i>speech-to-text</i>)
Taxa de amostragem	número de amostras do sinal por segundo, em hertz (Hz); 16 kHz é o padrão mais comum em ASR
MFCC	<i>Mel-Frequency Cepstral Coefficients</i> : coeficientes cepstrais na escala mel, atributo acústico clássico
Espectrograma log-mel	representação tempo-frequência com energia em bandas da escala mel, em escala logarítmica; entrada típica dos modelos neurais
Modelo acústico	componente que estima a probabilidade do sinal observado dado uma unidade linguística (fonema, letra ou subpalavra)
Léxico	dicionário de pronúncias que liga palavras a sequências de fonemas
Modelo de língua (LM)	componente que estima a probabilidade de uma sequência de palavras; LM vem de <i>language model</i> , que também se traduz por modelo de linguagem (seção 4.2)
HMM	<i>Hidden Markov Model</i> : modelo oculto de Markov, base do ASR estatístico clássico
OOV	<i>Out-of-vocabulary</i> : palavra fora do vocabulário do sistema
WER	<i>Word Error Rate</i> : taxa de erro de palavras, $(S + D + I) / N$
CER e SER	<i>Character Error Rate</i> : taxa de erro de caracteres; SER, <i>Sentence Error Rate</i> : taxa de erro de sentenças
RTF	<i>Real-Time Factor</i> : tempo de processamento dividido pela duração do áudio
Parâmetros	números internos de um modelo, ajustados no treinamento (também chamados pesos)
DNN	<i>Deep Neural Network</i> : rede neural profunda, com muitas camadas

ESQUEMA O mapa do capítulo

- **Seção 2.** O que é ASR e o que não é: transcrição, identificação de locutor, compreensão e síntese.
- **Seção 3.** A fala como sinal: amostragem, quantização, formatos e atributos acústicos (espectrograma, mel, MFCC).
- **Seção 4.** Arquitetura clássica: equação fundamental, modelo acústico, léxico, modelo de língua e decodificador; HMM-GMM e híbrido DNN-HMM; noções de aprendizado de máquina e de redes neurais.
- **Seção 5.** Avaliação: WER, CER, SER, confiança, latência e fator de tempo real.
- **Seção 6.** Fatores que afetam o desempenho: acústicos, linguísticos, do falante e de domínio.

1. INTRODUÇÃO

Reconhecimento automático de fala é a tarefa de converter um sinal de voz em uma sequência de palavras escritas. A definição parece simples, mas reúne três problemas de natureza distinta. **O primeiro** é físico: a voz chega ao sistema como variação de pressão do ar captada por um microfone, e precisa ser digitalizada e descrita por números. **O segundo** é linguístico: os mesmos sons podem corresponder a palavras diferentes ("concerto" e "conserto"), e a fala real é cheia de hesitações, repetições e frases interrompidas. **O terceiro** é estatístico: o sistema precisa escolher, entre milhões de sequências possíveis, aquela que mais provavelmente foi dita.

Para o cargo de Analista Legislativo na especialidade de Registro e Redação, esse conhecimento tem aplicação imediata. A Câmara dos Deputados utiliza transcrição automática integrada ao seu sistema interno de registro, e a qualidade do texto produzido pela máquina determina o volume e a natureza do trabalho de revisão humana. Saber por que o sistema erra, onde erra com mais frequência e como esse erro é medido é o que permite ao analista usar a ferramenta com critério.

Este capítulo trata da base comum a todos os sistemas. O Capítulo 2 examina os modelos que aprendem de ponta a ponta (*end-to-end*), como o Whisper, modelo de reconhecimento de fala lançado pela OpenAI em 2022, e os grandes modelos de linguagem que recebem áudio diretamente; o Capítulo 3 trata da diarização, da pontuação automática e da transcrição

em tempo real; o Capítulo 4 discute o português brasileiro e as ferramentas comerciais.

1.1. Como ler os itens desta disciplina

Os itens sobre reconhecimento de fala costumam combinar uma definição técnica com uma afirmação sobre o alcance da tecnologia. O erro mais comum é atribuir a um componente a função de outro (dizer que o modelo de língua converte o som em fonemas, quando isso é papel do modelo acústico) ou exagerar a capacidade do sistema (afirmar que a transcrição automática dispensa revisão, ou que um modelo é imune a ruído). Convém desconfiar de absolutos como "elimina", "garante", "independentemente" e "em qualquer condição", que raramente descrevem bem uma tecnologia probabilística.

■ QUADRO 1.2
COMO LER UM ITEM SOBRE ASR

1	Identificar o componente: perguntar de qual parte do sistema o item fala (sinal, atributos, modelo acústico, léxico, modelo de língua, decodificador, pós-processamento) e conferir se a função atribuída é a correta.
2	Conferir a métrica: verificar se a fórmula, o sentido (quanto menor a WER, melhor) e o alcance da métrica estão corretos.
3	Desconfiar do absoluto: sistemas de ASR são probabilísticos; afirmações de acerto total, imunidade a ruído ou dispensa de revisão humana tendem a estar erradas.

2. O QUE É RECONHECIMENTO AUTOMÁTICO DE FALA

2.1. Definição e tarefas vizinhas

O ASR (*Automatic Speech Recognition*, reconhecimento automático de fala) também é chamado de STT (*speech-to-text*, fala para texto) ou, no uso corrente, de transcrição automática. Sua entrada é áudio; sua saída é texto. O sistema não precisa "entender" o que foi dito para transcrever: um bom reconhecedor pode transcrever corretamente uma frase sem nenhum sentido, e um reconhecedor ruim pode errar uma frase simples. A compreensão do conteúdo é tarefa de outra área, a compreensão de linguagem natural (NLU, *Natural Language Understanding*).

Várias tecnologias de fala costumam ser confundidas com o ASR. A síntese de fala (TTS, *text-to-speech*) faz o caminho inverso, de texto para voz. A identificação e a verificação de locutor (*speaker identification* e *speaker verification*) não perguntam o que foi dito, mas quem disse: a primeira escolhe uma pessoa entre várias conhecidas, a segunda confirma se a voz pertence a quem alega ser, como em sistemas de autenticação biométrica. A diarização (Capítulo 3) pergunta "quem falou quando", separando os trechos de cada orador sem necessariamente saber seus nomes. A detecção de atividade de voz (VAD, *Voice Activity Detection*) apenas separa trechos com fala de trechos de silêncio ou ruído. A identificação de idioma (LID, *Language Identification*) reconhece qual língua está sendo falada.

■ QUADRO 1.3

TECNOLOGIAS DE FALA E A PERGUNTA QUE CADA UMA RESPONDE

Tecnologia	Pergunta e saída
ASR (STT)	O que foi dito? Saída: texto
TTS	Como dizer este texto? Saída: áudio sintetizado
VAD	Há fala neste trecho? Saída: marcação de início e fim de fala
Identificação de locutor	Quem, entre pessoas conhecidas, está falando? Saída: identidade
Verificação de locutor	Esta voz é de quem alega ser? Saída: aceita ou rejeita
Diarização	Quem falou quando? Saída: segmentos rotulados (Locutor 1, Locutor 2...)
LID	Em que língua se fala? Saída: código do idioma
NLU	O que o texto significa? Saída: intenção, entidades, sentido

2.2. Breve histórico

Os primeiros reconhecedores, nos anos 1950, identificavam dígitos falados por um único locutor. Nas décadas de 1970 e 1980, a abordagem estatística baseada em modelos ocultos de Markov (HMM) consolidou-se e dominou o campo por cerca de trinta anos, combinada a modelos de mistura de gaussianas (GMM) para descrever os sons. A partir de 2010, as redes neurais profundas (seção 4.4) substituíram as gaussianas no modelo acústico (sistemas híbridos DNN-HMM) e reduziram fortemente os erros. Ao longo da década de 2010, sobretudo na segunda metade, ganharam

espaço os modelos de ponta a ponta (*end-to-end*), que aprendem diretamente a relação entre áudio e texto (suas bases, como o CTC, de 2006, são anteriores), e, a partir de 2020, os modelos treinados com volumes muito grandes de dados, como o wav2vec 2.0 (aprendizado auto-supervisionado) e o Whisper (supervisão fraca em 680 mil horas de áudio), formas de aprendizado explicadas na seção 4.4. Os grandes modelos de linguagem capazes de receber áudio diretamente representam a etapa mais recente.

■ FIGURA 1.4

DO ESTATÍSTICO AO NEURAL: A EVOLUÇÃO DO ASR

Anos 1950-1960	Reconhecedores de dígitos e palavras isoladas: vocabulário mínimo, um único locutor
Anos 1970-2000	HMM-GMM: modelo estatístico com modelo acústico, léxico e modelo de língua separados
Década de 2010	Híbrido DNN-HMM: redes neurais profundas no modelo acústico, mantida a estrutura do HMM
Década de 2010 (sobretudo a segunda metade)	End-to-end: CTC, RNN-T e modelos de atenção com codificador (<i>encoder</i>) e <i>decoder</i> aprendem áudio-texto diretamente
A partir de 2020	Grandes modelos: auto-supervisão (wav2vec 2.0), supervisão fraca (Whisper), LLMs multimodais com áudio nativo

Esquema simplificado. As fases se sobrepõem: sistemas híbridos continuaram em uso comercial depois do surgimento dos modelos end-to-end. CTC (Connectionist Temporal Classification, classificação temporal conexionista), RNN-T (Recurrent Neural Network Transducer, transdutor de rede neural recorrente) e os modelos com codificador (encoder), que lê o áudio, e decoder, que gera o texto, são as famílias de modelos de ponta a ponta (Capítulo 2, seção 3). LLMs multimodais são grandes modelos de linguagem que recebem outras modalidades além do texto, como áudio e imagem (Capítulo 2, seção 6).

🔪 ONDE O ITEM PODE ERRAR

Troca de tarefa: "a verificação de locutor converte a fala em texto e identifica o orador" (errado: a verificação confirma a identidade alegada; quem converte fala em texto é o ASR). Troca de sentido: "TTS é a sigla do reconhecimento de fala" (errado: TTS é síntese, do texto para a voz).

Exagero: "para transcrever corretamente, o sistema de ASR precisa compreender o significado da frase" (errado: o reconhecimento é estatístico; o contexto linguístico ajuda, mas o sistema não precisa compreender o conteúdo no sentido humano).

PALAVRAS-CHAVE**ASR / STT** · conversão de fala em texto**Verificação de locutor** · confirmar se a voz é de quem alega ser (biometria)**Diarização** · quem falou quando, sem precisar saber o nome**VAD** · separa fala de silêncio e ruído

3. A FALA COMO SINAL

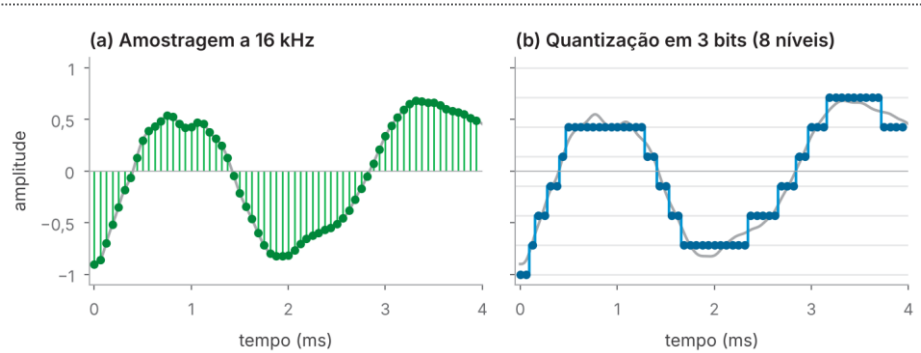
3.1. Do som ao sinal digital

A voz é uma onda de pressão sonora. O microfone converte essa onda em sinal elétrico contínuo (analogico), e um conversor analógico-digital transforma esse sinal em números por meio de duas operações. A **amostragem** mede o valor do sinal a intervalos regulares; o número de medições por segundo é a taxa de amostragem, expressa em hertz. A **quantização** representa cada amostra com um número finito de bits, o que define a resolução da medida (16 bits por amostra é o padrão mais comum).

O teorema da amostragem (Nyquist-Shannon) estabelece que, para representar sem perda as frequências até determinado valor, a taxa de amostragem deve ser de pelo menos o dobro desse valor (a rigor, mais que o dobro; na prática, usa-se uma margem). A telefonia tradicional usa 8 kHz, o que preserva frequências até 4 kHz e explica por que a voz ao telefone soa "abafada". O padrão mais usado em reconhecimento de fala é 16 kHz, que preserva frequências até 8 kHz, faixa onde se concentra a maior parte da informação útil da fala. Gravações de estúdio e de música usam 44,1 kHz ou 48 kHz; para o reconhecimento, esse áudio costuma ser reamostrado para 16 kHz, em geral sem prejuízo relevante.

■ FIGURA 1.5

AMOSTRAGEM E QUANTIZAÇÃO DE UM TRECHO DE VOZ



Trecho de 4 ms do sinal de uma frase produzida por sintetizador de voz (dados reais, não ilustrativos). A quantização em 3 bits serve apenas à visualização; o padrão mais comum é de 16 bits por amostra (65.536 níveis).

■ QUADRO 1.6

TAXAS DE AMOSTRAGEM E USOS

Taxa	Frequência máxima representada	Uso típico
8 kHz	4 kHz	telefonia, centrais de atendimento
16 kHz	8 kHz	padrão do ASR; entrada do Whisper e da maioria dos modelos
44,1 kHz e 48 kHz	22,05 kHz e 24 kHz	música, vídeo, gravação de estúdio

O sinal digitalizado pode ser armazenado sem compressão (formato WAV com PCM, *Pulse-Code Modulation*, modulação por código de pulsos, isto é, as amostras gravadas tal como medidas) ou com compressão. A compressão sem perdas (formato FLAC) reduz o tamanho do arquivo e preserva o sinal original; a compressão com perdas (formatos MP3, AAC e Opus) descarta informação considerada pouco perceptível ao ouvido humano. Para o ouvinte, a diferença pode ser imperceptível; para o reconhecedor, taxas de compressão agressivas podem remover detalhes acústicos úteis e aumentar os erros. Por isso, em acervos que serão transcritos e preservados, recomenda-se manter uma cópia de alta qualidade do áudio original.

Dois outros parâmetros do sinal importam. O número de canais distingue áudio mono (um canal) de estéreo ou multicanal; em plenários, cada microfone pode gerar um canal separado, o que facilita identificar quem fala. A relação sinal-ruído (SNR, *signal-to-noise ratio*), medida em decibéis, indica quanto a voz se destaca do ruído de fundo; quanto menor a SNR, mais difícil o reconhecimento.

3.2. Noções de fonética úteis ao ASR

A unidade mínima que distingue significado numa língua é o **fonema**: em português, /p/ e /b/ (entre barras, indica-se o fonema, e não a letra) são fonemas distintos porque "pato" e "bato" são palavras diferentes. A realização concreta de um fonema varia conforme o contexto e o falante, e essas variantes são chamadas **alofones**: o /t/ de "tia" é pronunciado como "tchia" em boa parte do Brasil e como "tia" em outras regiões. Os sons também se influenciam mutuamente na fala contínua, fenômeno chamado **coarticulação**: a articulação de um som antecipa a do seguinte, de modo que não há fronteiras nítidas entre fonemas, nem, muitas vezes, entre palavras.

A **prosódia** reúne os aspectos de entonação, ritmo, duração e intensidade. Ela distingue uma pergunta de uma afirmação ("Vossa Excelência votou." e "Vossa Excelência votou?") e marca fronteiras de frases, o que torna a prosódia relevante para a pontuação automática (Capítulo 3). Por fim, a relação entre letras e sons não é biunívoca: a letra "x" tem pelo menos quatro valores em português ("xícara", "exame", "táxi", "próximo"), e sons iguais podem ser escritos de formas diferentes ("cessão", "seção", "sessão"). Essas ambiguidades são resolvidas, em grande parte, pelo modelo de língua.

3.3. Atributos acústicos: do sinal ao espectrograma

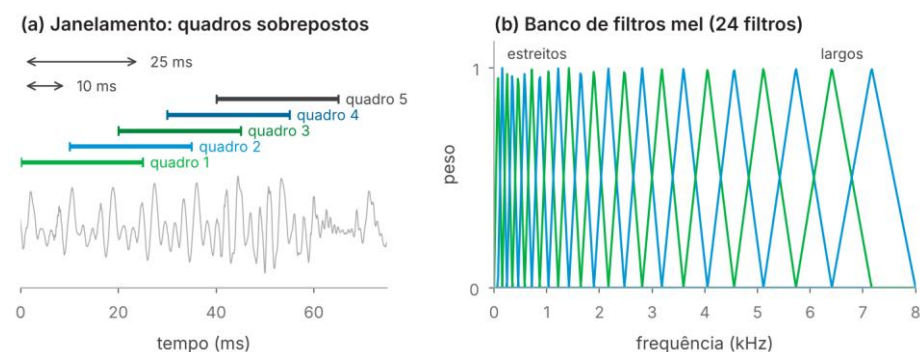
O sistema não trabalha diretamente com as amostras do sinal, que são muitas (16 mil por segundo a 16 kHz) e pouco informativas isoladamente. O sinal é dividido em **quadros** (*frames*) curtos (quadro, aqui, é um trecho do sinal, sem relação com os quadros deste livro), tipicamente janelas de 25 milissegundos que avançam de 10 em 10 milissegundos, de modo que quadros vizinhos se sobrepõem. Dentro de um intervalo tão curto, a fala pode ser tratada como aproximadamente estável. Cada quadro passa por

uma transformada de Fourier, que decompõe o sinal nas frequências que o compõem. O resultado, quadro a quadro, é o **espectrograma**: uma imagem com o tempo no eixo horizontal, a frequência no eixo vertical e a energia representada pela intensidade.

A percepção humana de altura não é linear: distinguimos bem diferenças entre 200 Hz e 300 Hz, mas mal entre 7.000 Hz e 7.100 Hz. A **escala mel** imita essa percepção, com bandas estreitas nas frequências baixas e largas nas altas. Agrupando a energia do espectro em bandas mel e aplicando o logaritmo (que também imita a percepção de intensidade), obtém-se o **espectrograma log-mel**, entrada padrão dos modelos neurais modernos. Cada banda é obtida por um filtro, e o conjunto deles forma o **banco de filtros mel**; na literatura, as bandas também são chamadas canais mel, sem relação com os canais de áudio da seção 3.1. Cada quadro passa a ser descrito por uma lista de números, um por banda, chamada **vetor** (seção 4.4). O Whisper, por exemplo, recebe espectrogramas log-mel de 80 bandas mel nas primeiras versões e de 128 bandas mel na versão large-v3 (Capítulo 2, seção 5.3).

■ FIGURA 1.7

JANELAMENTO E BANCO DE FILTROS MEL



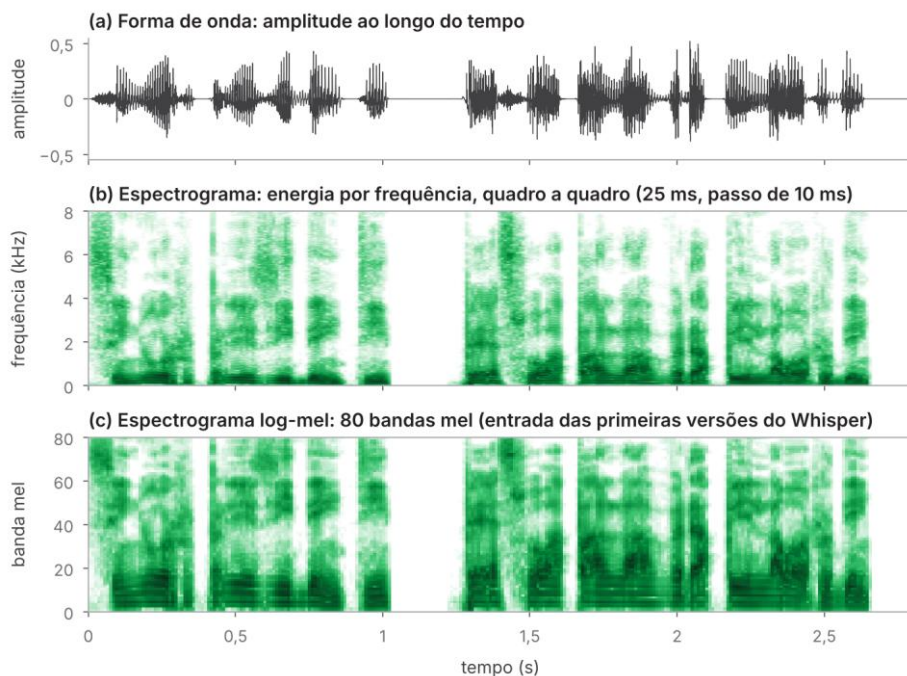
Em (a), trecho real de 75 ms: cada quadro de 25 ms começa 10 ms depois do anterior. Em (b), 24 filtros (24 bandas mel), para facilitar a leitura; o Whisper usa 80 ou 128 bandas.

Os **MFCC** (*Mel-Frequency Cepstral Coefficients*, coeficientes cepstrais na frequência mel) foram o atributo clássico da era HMM-GMM. Obtém-se aplicando ao espectro log-mel uma transformada do cosseno, que comprime a informação em poucos coeficientes pouco correlacionados (em geral 13), aos quais se acrescentam suas variações no tempo (deltas e delta-

deltas), totalizando 39 valores por quadro. Pouco correlacionados quer dizer que repetem pouca informação entre si, o que convinha aos modelos de gaussianas da época, que tratavam cada coeficiente de forma independente. O nome cepstral vem de *cepstrum*, palavra formada pela inversão das quatro primeiras letras de *spectrum* (espectro), porque o resultado é obtido por uma transformação aplicada ao próprio espectro. Com as redes neurais, que lidam bem com dados correlacionados, o espectrograma log-mel passou a ser preferido, e modelos como o wav2vec 2.0 chegam a aprender seus próprios atributos diretamente da forma de onda.

■ FIGURA 1.8

A MESMA FRASE EM TRÊS REPRESENTAÇÕES



Frase "Senhor Presidente, peço a palavra pela ordem.", produzida por sintetizador de voz (eSpeak NG com voz MBROLA) e analisada pelo autor. Cores mais escuras indicam mais energia.

- **Fatores de erro:** ruído, reverberação, distância do microfone, sobreposição, compressão, fala espontânea, sotaque, grupo de falantes e vocabulário de domínio. Robustez vem de dados diversos, aumento de dados (SpecAugment), pré-processamento e adaptação; nenhuma técnica torna o sistema imune.

QUESTÕES DE CONCURSO COMENTADAS

► **Questão 1** (Cebraspe/Câmara dos Deputados/Policial Legislativo Federal/2026)

Em relação a Android, *malware* e IA generativa, julgue o item subsequente.

Em sistemas de IA generativa baseados em modelos de linguagem, o processo de geração de saída envolve a predição iterativa autorregressiva de *tokens* com base em distribuições de probabilidade condicionadas ao contexto de entrada e ao histórico de *tokens* previamente gerados, o que permite a construção incremental de sequências coerentes.

() Certo () Errado

Gabarito: Certo

Comentários: Certo. O item descreve como os modelos generativos de linguagem produzem a saída: um *token* de cada vez (o *token* é a unidade que o modelo lê ou gera, em geral um pedaço de palavra; ver o boxe da seção 4.2), cada um escolhido a partir de uma **distribuição de probabilidade**, como a que a função softmax fornece na última camada da rede (seção 4.4), condicionada à entrada e aos *tokens* já gerados. Esse encadeamento, em que cada saída depende das anteriores, é a geração **autorregressiva**: o texto é construído de forma incremental, e a dependência do que já foi escrito é o que lhe dá coerência.

O mesmo mecanismo explica o principal risco desses modelos: como cada *token* é escolhido por ser provável depois dos anteriores, o modelo pode escrever, passo a passo, texto fluente que não foi dito ou que não é verdadeiro (alucinação; seção 5.3). O Whisper transcreve da mesma forma, condicionado ao áudio (Capítulo 2, seção 3.3), e os grandes modelos de linguagem seguem a mesma lógica (Capítulo 2, seção 6.1). Ver o boxe sobre IA generativa da seção 4.4.

QUESTÕES INÉDITAS COMENTADAS

► Questão 2 (Inédita, nível fácil)

Acerca dos conceitos básicos de reconhecimento automático de fala, julgue o item a seguir. A verificação de locutor é a tarefa de converter a fala em texto, identificando, ao mesmo tempo, quem é o orador.

() Certo () Errado

Gabarito: Errado

Comentários: A conversão de fala em texto é o reconhecimento automático de fala (ASR ou STT). A verificação de locutor responde a outra pergunta: se a voz pertence a quem alega ser, como em sistemas de autenticação biométrica. O item mistura duas tarefas distintas e ainda atribui à verificação a função da identificação de locutor.

Convém fixar o quadro da seção 2.1: ASR pergunta o que foi dito; identificação, quem fala entre pessoas conhecidas; verificação, se é quem diz ser; diarização, quem falou quando.

► Questão 3 (Inédita, nível fácil)

Acerca dos conceitos básicos de reconhecimento automático de fala, julgue o item a seguir. A diarização de falantes busca responder à pergunta "quem falou quando", segmentando o áudio por orador, o que não exige, necessariamente, que o sistema conheça os nomes das pessoas que falaram.

() Certo () Errado

Gabarito: Certo

Comentários: A diarização separa os trechos de cada orador e os rotula (Locutor 1, Locutor 2), sem exigir identificação nominal. Associar os rótulos aos nomes é etapa adicional, que pode usar identificação de locutor, roteiros da sessão ou trabalho humano. Ver a seção 2.1 e o Capítulo 3.

► Questão 4 (Inédita, nível médio)

Em relação à digitalização do sinal de voz, julgue o item a seguir.

Para representar fielmente as frequências da fala de até 8 kHz, é suficiente uma taxa de amostragem de 8 kHz, razão pela qual esse valor é o padrão adotado pela maioria dos sistemas modernos de reconhecimento de fala.

() Certo () Errado

Gabarito: Errado

Comentários: Pelo teorema da amostragem (Nyquist-Shannon), a taxa deve ser de pelo menos o dobro da maior frequência que se quer representar. Para preservar até 8 kHz, são necessários 16 kHz, que é justamente o padrão mais comum em ASR. A taxa de 8 kHz é a da telefonia e só preserva frequências até 4 kHz.

O item tem dois erros encadeados: a regra de Nyquist invertida e o padrão do ASR trocado pelo da telefonia. Ver a seção 3.1.