UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Sobre a dependência estatística entre MFCCs e a combinação de energia e frequência fundamental

Vitor Magno de Oliveira Santos Bezerra

Cidade Universitária Prof. José Aloísio de Campos
São Cristóvão-SE, Brasil

04/08/2026

FICHA CATALOGRÁFICA ELABORADA PELO SIBIUFS
UNIVERSIDADE FEDERAL DE SERGIPE

Bezerra, Vitor Magno de Oliveira Santos
B574s Sobre a dependência estatística entre MFCCs e a combinação de energia e frequência fundamental / Vitor Magno de Oliveira Santos Bezerra ; orientador Jugurta Rosa Montalvão Filho. – São Cristóvão, SE, 2026.
53 f. ; il.

Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal de Sergipe, 2026.

1. Engenharia elétrica. 2. Fala. 3. Estatística. 4. Entropia. 5. Audição (Fisiologia) I. Montalvão Filho, Jugurta Rosa, orient. II. Título.

CDU 621.3:534.88

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA-PROEE

TERMO DE APROVAÇÃO

“Sobre a dependência estatística entre MFCCs e a combinação de energia e frequência fundamental "

Discente:

Vitor Magno de Oliveira Santos Bezerra

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Sergipe, como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica.

Aprovada pela banca examinadora composta por:

Documento assinado digitalmente
gov.br
EDUARDO OLIVEIRA FREIRE
Data: 01/09/2026 11:45:26-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Eduardo Oliveira Freire (PROEE/UFS)
Presidente

Documento assinado digitalmente
gov.br
JANIO COUTINHO CANUTO
Data: 01/09/2026 11:36:03-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Janio Coutinho Canuto (DCOMP/UFS)
Examinador Externo

Documento assinado digitalmente
gov.br
ELYSON ADAN NUNES CARVALHO
Data: 31/08/2026 10:09:09-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Elyson Ádan Nunes Carvalho (PROEE/UFS)
Examinador Externo

Documento assinado digitalmente
gov.br
VITOR MAGNO DE OLIVEIRA SANTOS BEZERRA
Data: 15/09/2026 21:23:37-0300
Verifique em https://validar.iti.gov.br

Vitor Magno de Oliveira Santos Bezerra
Discente

Cidade Universitária “Prof. José Aloísio de Campos”, 20 de agosto de 2026.

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Vitor Magno de Oliveira Santos Bezerra

Sobre a dependência estatística entre MFCCs e a combinação de energia e frequência fundamental

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica – PROEE, da Universidade Federal de Sergipe, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Jugurta Rosa Montalvão Filho

Cidade Universitária Prof. José Aloísio de Campos
São Cristóvão-SE, Brasil
04/08/2026

Resumo

Os coeficientes cepstrais de frequência em escala mel (MFCCs) constituem uma característica importante no processamento de fala. Este estudo investiga a suposição de que os valores de MFCCs são independentes dos valores da combinação de energia e frequência fundamental, que carregam informações prosódicas. Essa investigação é feita com o desenvolvimento de um teste de hipótese nula para avaliar a significância estatística da hipótese de independência entre as duas fontes de informação. Os resultados obtidos demonstram ser estatisticamente implausível que os MFCCs sejam independentes da combinação de energia e frequência fundamental. Além disso, a reprodução de um sistema de reconhecimento de orador publicado na literatura revela uma evidência prática dessa implausibilidade, à medida que os resultados sugerem que esses dois tipos de características não se complementam, em termos de ganhos de desempenho.

Palavras-chave: MFCCs; Hipótese Nula; Entropia; Processamento de Fala; Reconhecimento de Orador.

Abstract

Mel-frequency cepstral coefficients (MFCCs) are a key feature in speech processing. This study investigates the assumption that MFCC values are independent of the combination of energy and fundamental frequency, which carry prosodic information. This investigation is conducted by developing a null hypothesis test to assess the statistical significance of the independence hypothesis between these two sources of information. The results demonstrate that it is statistically implausible for MFCCs to be independent of the combination of energy and fundamental frequency. Furthermore, replicating a speaker recognition system from the literature provides practical evidence of this implausibility, as the results suggest that these two types of features do not complement each other in terms of performance gains.

Key-words: MFCCs; Null Hypothesis; Entropy; Speech Processing; Speaker Recognition.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

DCT	Transformada Discreta de Cosseno
DET	Compromisso entre Erro de Detecção
DFT	Transformada Discreta de Fourier
DI	Dimensão Intrínseca
EER	"Equal Error Rate"
EM	Expectativa-Maximização
F0	Frequência Fundamental
F_s	Frequência de Amostragem
FAR	Taxa de Falsa Aceitação
FRR	Taxa de Falsa Rejeição
GMM	Modelo de Mistura Gaussiana
LLR	Logaritmo da Razão de Verossimilhança
MFCC	Coeficiente Cepstral em Frequência Mel
NCCF	Função de Autocorrelação Cruzada Normalizada
RAPT	"Robust Algorithm for Pitch Tracking"
SNR_Q	Relação Sinal-Ruído de Quantização
UBM	Modelo de Fundo Universal

Sumário

1 Introdução

A fala é o principal método de comunicação da humanidade (FLANAGAN, 1972). Ela deriva, simultaneamente, da habilidade humana de produzir uma vasta gama de sons e da capacidade de associá-los aos símbolos e regras estruturais de um idioma (WAYLAND, 2018). Estima-se que os milhares de idiomas existentes no mundo sejam formados a partir de combinações de apenas algumas centenas de padrões sonoros distintos (ASHBY; MAIDMENT, 2005). Essa repetição de padrões entre línguas diferentes ocorre porque todas utilizam o mesmo sistema biológico gerador: o aparelho fonador. Especificamente, cada padrão sonoro está associado a variações na pronúncia que decorrem, essencialmente, de diferentes configurações de um mesmo elemento, o trato vocal.

Dessa forma, as diferentes configurações de formato do trato vocal determinam o conteúdo linguístico da mensagem. Paralelamente, essas mesmas configurações transmitem pistas sobre a identidade do orador, como o timbre da voz durante a emissão de uma vogal. Os sons vocálicos são produzidos por vibrações que ressoam no interior do trato vocal. Essa ressonância não apenas define a sonoridade da vogal em si, mas também carrega assinaturas acústicas das dimensões físicas do trato vocal. Como essas dimensões variam entre indivíduos, o timbre resultante torna-se único. É essa assinatura acústica singular que nos permite, por exemplo, reconhecer instantaneamente a voz de um conhecido durante uma chamada telefônica (LANCKER; CANTER, 1982).

Além dos determinantes de natureza fisiológica, a fala carrega uma forte dimensão comportamental. A enunciação das pronúncias não ocorre de forma isolada, mas com um encadeamento contínuo para expressar intenções e emoções. Esse encadeamento é modulado, em grande parte, pelos padrões de variação na frequência fundamental (F0) de vibração das cordas vocais (os pulsos glotais), dando origem à prosódia, que é comumente chamada de ‘melodia da fala’. Essa modulação que confere coesão de sentido ao discurso, permitindo que uma mesma sequência de palavras mude completamente de significado se dita como uma pergunta, uma ordem ou uma ironia. Uma fala emitida de maneira estritamente monotônica soa artificial e carece de naturalidade (FANT, 1960). A relevância desse componente comportamental é tão expressiva que imitadores profissionais não replicam apenas o timbre de uma pessoa, mas também consideram suas variações de tonalidade (ASHOUR; GATH, 1999; KITAMURA, 2008).

Dessa forma, a identidade do orador pode ser aferida por meio de elementos de natureza tanto fisiológica quanto comportamental (KUWABARA; SAGISAK, 1995). Na literatura, essa distinção encontra forte fundamentação no modelo fonte-filtro da produção da fala, que tradicionalmente assume que essas duas componentes operam de maneira independente (FANT, 1960; TITZE, 2008). Sob essa premissa de independência, os pulsos glotais atuam como uma

fonte acústica autônoma de um sinal de som que é posteriormente modificado pelas propriedades de filtragem do trato vocal. A partir desse pressuposto, uma vertente de pesquisa concentrou-se no desenvolvimento de sistemas ligados ao comportamento do orador por meio das características prosódicas (ATAL, 1972; SÖNMEZ et al., 1997; ADAMI et al., 2003; SHRIBERG et al., 2005; DEHAK; DUMOUCHEL; KENNY, 2007). Enquanto isso, outra vertente se dedicou ao estudo de características espectrais para mapear os traços anatômicos do trato vocal (ATAL, 1974; SOONG et al., 1985; REYNOLDS; QUATIERI; DUNN, 2000; DEHAK et al., 2011; SNYDER et al., 2018). Sob a ótica de independência entre as características, o desempenho de ambos os sistemas deveria ser complementar, dado que cada sistema foca em dimensões teoricamente isoladas do sinal de fala.

Contudo, resultados experimentais indicam que a combinação direta de características prosódicas com os coeficientes cepstrais nas frequências mel (MFCCs) resulta apenas em melhorias marginais de desempenho, com destaque especial para o trabalho de Adami (2007) que constitui a principal referência do presente trabalho. Como os MFCCs são características espectrais incorporando propriedades perceptuais do sistema auditivo humano (DAVIS; MERMELSTEIN, 1980), seria esperado que a sua fusão com dados prosódicos gerasse um ganho expressivo de informação complementar. A ausência desse salto de desempenho levanta duas hipóteses fundamentais, as quais podem coexistir: o pressuposto de independência entre fonte e filtro pode não se sustentar na prática, e/ou os MFCCs já estão capturando intrinsecamente ambos os tipos de informação (fisiológica e comportamental).

Questionamentos sobre a natureza ou características da fala podem parecer anacrônicos em relação ao paradigma atual de processamento de fala. Nesse paradigma, o estado da arte atingiu resultados impressionantes usando modelos de rede neural "end-to-end" que operam diretamente com sinais de falas brutos ou espectrogramas, sem a necessidade de pré-processamento (MEHRISH et al., 2023). Assim, as informações nesses sinais são capturadas por representações aprendidas pelo próprio modelo. No entanto, tais representações são internas à arquitetura, a qual funciona como uma "caixa preta", dificultando a interpretabilidade de suas decisões (CHOWDHURY; DURRANI; ALI, 2024). Além disso, esse abordagem restringe-se a cenários que dispõem de volumes massivos de dados e elevado poder computacional (GLASMACHERS, 2017; THOMPSON et al., 2020). Em contextos distintos, portanto, os sistemas ainda precisam recorrer ao uso de características extraídas manualmente, como os MFCCs.

Um exemplo de cenário intrinsecamente com baixo volume de dados é a área médica (MOLLAIE et al., 2026; AHSAN; LUNA; SIDDIQUE, 2022), na qual os MFCCs são utilizados como biomarcadores para monitoramento de saúde por meio da fala, auxiliando na detecção de patologias como as doenças de Parkinson e Alzheimer, além da depressão (TRACEY et al., 2023; HOSSAIN; TRAINI; AMENTA, 2026). Ademais, os MFCCs são amplamente empregados na detecção de palavras-chave (LÓPEZ-ESPEJO et al., 2021; GARAI; SAMUI, 2025). Uma

aplicação com baixo recurso computacional que, além de ser usada independentemente, auxilia na implementação de modelos "end-to-end": um sistema leve baseado em MFCCs monitora constantemente o aúdio de forma nativa e econômica, ativando o modelo neural mais robusto somente quando a palavra-chave é detectada (YANG, 2026; SHAN et al., 2020). O reconhecimento de palavras é uma atividade que tradicionalmente usa os MFCCs desconsiderando qualquer dependência temporal (NASSIF et al., 2019; ABDUL; AL-TALABANI, 2022). Supõe-se, contudo, que a modelagem dessa dependência temporal possa viabilizar a recuperação da prosódia, permitindo o desenvolvimento de sistemas de diagnóstico e monitoramento mais robustos a partir da combinação de ambos os conjuntos de informação.

O estudo da integração de MFCCs com características prosódicas teve destaque no estado da arte do reconhecimento de orador nos anos 2000 (SHRIBERG et al., 2005; DEHAK; DUMOUCHEL; KENNY, 2007; ADAMI, 2007). Contudo, essa vertente foi ofuscada por não oferecer saltos qualitativos de desempenho, enquanto a abordagem dos "i-vectors" introduzida em Dehak et al. (2011), mostrou-se significativamente mais promissora, e posteriormente, os "x-vectors" adaptou essa abordagem para as redes neurais profundas (SNYDER et al., 2018). Apesar desse ofuscamento, a combinação dos dois conjunto permanece útil em situações específicas, como um sistema de reconhecimento para oradores com disartria (SALIM; SHAHNAWAZUDDIN; AHMAD, 2023), ou na investigação do comportamento dos sistemas, como em falhas de reconhecimento quando os MFCCs estão associados a F0 distantes da F0 média do orador (HAUTAMÄKI; KINNUNEN, 2020). Na literatura, sabe-se, por exemplo, que os MFCCs carregam informações sobre a energia pontual do sinal, mas assume-se que eles sejam insensíveis às mudanças de F0 (ZHENG; LEE; CHING, 2007). No entanto, o trabalho em Milner, Shao e Darch (2005) indica que é possível recuperar a F0 com MFCCs através do uso de Modelos Ocultos de Markov.

Diante desse cenário, a presente dissertação propõe uma metodologia para investigar se há independência entre os dois conjuntos de características utilizados em Adami (2007) (MFCCs e energia + F0). Nessa metodologia, os vetores de características escalares extraídas em cada conjunto são quantizados por meio de Modelos de Mistura de Gaussianas (GMMs), mapeando seus valores contínuos em índices discretos. Como resultado, o sinal de fala passa a ser representado por duas sequências de símbolos discretos. Essa estratégia de discretização viabiliza o cálculo da dependência estatística entre as sequências por meio da entropia condicional para fontes discretas. A investigação apoia-se na premissa fundamental de que, se houver acoplamento de informação, a dependência estatística observada entre as duas características extraídas simultaneamente (alinhadas em sua sequência natural) será significativamente maior do que a dependência residual encontrada quando o alinhamento temporal entre as sequências é quebrado (como em distribuições permutadas aleatoriamente). A significância dessa discrepância estatística entre as sequências alinhadas e permutadas é estimada rigorosamente por um procedimento de teste de hipótese nula.

1.1 Objetivos

O objetivo geral desta dissertação é investigar a existência e a relevância de informações prosódicas intrínsecas aos MFCCs. Essa investigação é realizada por meio da implementação de testes de hipótese nula em um conjunto com 1600 falas em português brasileiro, cujos resultados servem como prova de conceito sobre o potencial de extrair informações prosódicas dos MFCCs. A validação estatística da dependência entre as características prosódicas e os MFCCs levará à implementação de um experimento de reconhecimento de orador, utilizando o mesmo conjunto de dados, cujo propósito é avaliar o impacto prático dessa dependência.

Os objetivos específicos que guiam este desenvolvimento compreendem:

- **Modelar** o espaço de características com símbolos por meio da quantização vetorial baseada em GMM;
- **Quantificar** a dependência entre as características via entropia condicional e determinar sua relevância estatística;
- **Avaliar** o desempenho das características analisadas em um experimento de reconhecimento de orador;
- **Analisar** o impacto prático da dependência entre as características no reconhecimento de orador.

1.2 Organização da Dissertação

O presente trabalho está estruturado em cinco capítulos, organizados da seguinte forma:

No **Capítulo 2**, dois procedimentos essenciais para esse trabalho, o processamento de fala e a estimação de entropia, são detalhados teoricamente e contextualizados por uma revisão bibliográfica.

No **Capítulo 3**, os procedimentos metodológicos adotados para a extração e quantização de características são descritos, a metodologia dos testes de hipóteses é delineada e o conjuntos de dados que foi utilizados é apresentado.

No **Capítulo 4**, os resultados dos testes de hipótese e o experimentos de reconhecimento de orador são apresentados e discutidos.

Por fim, no **Capítulo 5**, as considerações finais em relação a este trabalho são feitas, sintetizando as principais contribuições deste estudo e propondo diretrizes para trabalhos futuros.

2 Fundamentos Teóricos e Revisão da Literatura

O principal objetivo deste trabalho é investigar a independência entre os MFCCs e as características prosódicas. Essa investigação se baseia em testes de hipótese nula definidos no próximo capítulo, os quais seguem um procedimento padrão bem estabelecido na comunidade científica. Por ser uma metodologia consolidada, este capítulo não detalha o procedimento em si, focando apenas na fundamentação dos elementos que são utilizados para construir os experimentos. Assim, a Seção 2.1 é sobre características espectrais e prosódicas. A independência entre elas é avaliada com os elementos introduzidos na Seção 2.2 que é sobre entropia e o problema de sua estimação. Além da fundamentação, esse capítulo contém elementos de revisão bibliográfica, ao passo que também são mencionados trabalhos relevantes acerca dos conceitos abordados.

2.1 Sobre características espectrais e prosódicas

No processamento de fala, dois tipos de características se destacam. O primeiro deles é o das caraterísticas espetrais, que são tradicionalmente usadas para modelar apenas os efeitos segmentais do formato do trato vocal (CAMPBELL et al., 2003). Dentre estas, estão os MFCCs, que são o objeto de estudo deste trabalho. O segundo tipo de caraterística é o das características relacionadas a prosódia da fala, que são expressas, principalmente, através de três parâmetros acústicos: a energia, frequência fundamental (F0), e a duração (MARY; YEGNANARAYANA, 2008).

A respeito desses dois conjuntos de características, o modelo filtro-fonte linear postula que a ressonância do filtro do trato vocal (que é descrita pelas características espetrais) é independente do fluxo de ar periódico da fonte glotal (que é descrito pelas características prosódicas) (FANT, 1960; TITZE, 2008). Ao longo de muitas décadas, esse pressuposto foi implicitamente usado na literatura, de modo que o espectro e as variações temporais da fonte tenham sido tratados como independentes durante o desenvolvimento de métodos, e à medida que a combinação da informação desses dois conjuntos de descritores é conhecida por melhorar o desempenho em diversas aplicações envolvendo a fala (HASIJA et al., 2022; ALAMRI et al., 2026; VERMA et al., 2023; KURDI, 2026; MISSAOUI; LACHIRI, 2025).

Em contrapartida, para o caso específico dos MFCCs, alguns trabalhos voltados ao reconhecimento de orador parecem sugerir que esse conjunto de coeficientes (originalmente categorizados como descritores espectrais) inerentemente codificam a prosódia, o que contradiz a teoria clássica. Por exemplo, o trabalho em Adami (2007) revela que os resultados obtidos a partir

dos MFCCs são apenas marginalmente melhorados após a fusão com a informação prosódica. Além disso, outros estudos se mostraram capazes de recuperar informações prosódicas a partir dos MFCCs (KIM; PARK, 2020; MILNER; SHAO, 2006).

Essa observação foi o que motivou o presente estudo, que, como mencionado anteriormente, tem como objetivo fazer uma investigação sistemática da extensão dessa relação (entre MFCCs e prosódia). Dessa forma, o restante desta seção tem como objetivo trazer uma fundamentação teórica acerca dos MFCCs, e de como estes são calculados, e acerca das características prosódicas. Antes de adentrar no processo de extração dessas características, no entanto, introduz-se um aspecto comum ao processamento de fala, e que é usado ao longo deste trabalho: o processamento em tempo curto, que serve para lidar com a não-estacionaridade inerente ao processo de produção desses sinais (RABINER; SCHAFER, 2007).

2.1.1 Processamento em tempo curto

O sinal de fala é intrinsecamente não estacionário. Assim, suas propriedades estatísticas e espectros variam a cada fonema pronunciado (RABINER; SCHAFER, 2007). Consequentemente, o processamento de fala atua durante pequenos intervalos nos quais essas informações podem ser analisadas isoladamente.

Consideremos um sinal de fala $s[n]$ segmentado em N intervalos de fala com M amostras através da operação de janelamento (RABINER; SCHAFER, 2007). Assim, cada segmento está associado a duração de M/F_s segundos, em que F_s é a frequência de amostragem do sinal $s[n]$. As durações que são tipicamente utilizadas no processamento variam entre 10 ms a 30 ms. Além disso, alguns efeitos indesejados de falsa descontinuidade provocado pelas bordas impostas aos segmentos podem ser mitigados com o estabelecimento de uma taxa de sobreposição entre eles. Ao final desse processo, as M amostras do i-ésimo segmento de $s[n]$ são denotadas por $s^{(i)}[n]$.

Matematicamente, o isolamento das amostras de um segmento acontece pela multiplicação do sinal $s[n]$ com uma função de janela com M amostras (DENG; O'SHAUGHNESSY, 2003). Os diferentes segmentos da fala são obtidos através de deslocamentos da função de janela ao longo do sinal. Neste trabalho, duas funções de janela são utilizadas: a janela retangular e a janela de Hamming. A primeira é descrita pela equação

$$w_R[n] = \begin{cases} 1 & \text{para } 1 \leq n \leq M \\ 0 & \text{em outros casos} \end{cases}. \tag{2.1}$$

As extremidades dessa função variam bruscamente causando o fenômeno de vazamento espectral nos segmentos (OPPENHEIM; SCHAFER, 2010). Esse efeito pode ser mitigado com o uso da janela de Hamming que é descrita pela equação

$$w_H[n] = \begin{cases} 0.54 - 0.46\cos(2\pi(n-1)/(M-1)) & \text{para } 1 \leq n \leq M \\ 0 & \text{em outros casos} \end{cases}. \tag{2.2}$$

Para contexto do reconhecimento de fala, o processamento em tempo curto é um aspecto fundamental da etapa de extração de características, devido à necessidade de se levar em conta que os descritores dos sinais variam dinamicamente, sejam eles espectrais, prosódicos ou de outra natureza. Assim, durante esse processo, é comum que os atributos sejam extraídos de cada segmento de maneira independente (RABINER; SCHAFER, 2007). Ou seja, para cada atributo d que se deseja extrair de um sinal de áudio, obtém se uma sequência $\{d^{(i)}\}_i$, em que $d^{(i)}$ representa o valor desse atributo extraído do i-ésimo segmento. De fato, a abordagem do processamento em tempo curto é utilizada em todos os experimentos deste trabalho. Além disso, também é utilizada uma aproximação da primeira derivada das características para captar suas tendências dinâmicas de curto prazo (RABINER; SCHAFER, 2011). Assim, a variação da característica genérica $d^{(i)}$ associada ao i-ésimo segmento é calculada por meio da equação

$$\Delta d^{(i)} = \frac{\sum_{\tilde{\imath}=i-I}^{i+I} \tilde{\imath} d^{(\tilde{\imath})}}{\sum_{\tilde{\imath}=i-I}^{i+I} \tilde{\imath}^2}, \tag{2.3}$$

em que $\tilde{\imath}$ é um índice auxiliar para uma estimativa mais robusta da variação ao longo de $(2I + 1)$ segmentos centralizados em i. Essa abordagem é utilizada porque formulações mais simples baseadas em diferenças de primeira ordem tendem a ser excessivamente ruidosas (SOONG; ROSENBERG, 1988).

2.1.2 MFCCs

Por serem uma característica espectral, os MFCCs representam os diferentes formatos que o trato vocal assume durante a fala vozeada. Essa modelagem utiliza uma versão comprimida do espectro do sinal para captar com precisão os efeitos de ressonância da fala.

Os MFCCs foram propostos para emular o sistema auditivo humano (DAVIS; MERMELSTEIN, 1980), e possuem boa discriminação acústica por capturarem informações críticas relacionadas a padrões de fala e características vocais (VERMA et al., 2023). Além disso, os coeficientes são bem descorrelacionados, fornecendo uma versão comprimida do espectro que é bem conveniente para modelos estatísticos (TOLEDANO; FERNÁNDEZ-GALLEGO; LOZANO-DIEZ, 2018). Assim, apesar de terem sido proposta há décadas, eles ainda continuam sendo amplamente utilizados (ABDUL; AL-TALABANI, 2022).

O primeiro passo para o cálculo dos MFCCs é a aplicação do filtro de pré-ênfase ao sinal de fala $s[n]$ (ABDUL; AL-TALABANI, 2022). Esse filtro é um passa alta de primeira ordem que é utilizado para achatar o espectro de $s[n]$. O sinal resultante $s_p[n]$ é descrito pela equação

$$s_p[n] = s[n] - 0,97s[n-1]. \tag{2.4}$$

A próxima etapa é o janelamento para segmentar o sinal $s_p[n]$ (ABDUL; AL-TALABANI, 2022). Esse janelamento é realizado com janelas de Hamming para reduzir o vazamento espectral

, conforme explicado na Seção 2.1.1. Denotemos, então, por $s_p^{(i)}[n]$ o i-ésimo segmento do sinal s_p. Se o segmento tem duração de 20 ms, então $M = 220$ amostras quando a frequência de amostragem (F_s) tem 11,025 kHz.

Para o passo seguinte, o espectro de cada um dos segmentos é obtido com o uso da transforma discreta de Fourier (DFT). Cada segmento é preenchido com zeros para ficar com exatamente 256 amostras. Assim, o espectro de magnitude do i-ésimo segmento $F^{(i)}[k]$ é descrito pela equação

$$F^{(i)}[k] = \left| \sum_{n=1}^{256} s_p^{(i)}[n] \exp^{-j\frac{2\pi(n-1)(k-1)}{256}} \right|, \tag{2.5}$$

em que $k = 1, \ldots, 256$ são os índices associados a frequências.

A próxima etapa consiste em passar o espectro de magnitude por um banco de filtros triangulares. Na versão de MFCCs adotada neste trabalho, esse banco é composto por 24 filtros sobrepostos. Suas frequências centrais e de borda são determinadas por 26 frequências de corte $\{cf_h\}_{h=1}^{26}$ em Hz, as quais são definidas com espaçamento equidistante na escala mel. Assim, cf_1 e cf_{26} são usados para determinar a largura de banda completa na escala mel através da equação

$$f_{\text{mel}} = 2595 \log_{10}\left(1 + \frac{f_{\text{Hz}}}{700}\right), \tag{2.6}$$

em que f_{Hz} é a frequência em Hz e f_{mel} é a frequência em escala mel. A divisão dessa largura em 26 valores equidistantes permite o cálculo de cf_h através da equação

$$f_{\text{Hz}} = 700(10^{\frac{f_{\text{mel}}}{2595}} - 1). \tag{2.7}$$

Os 24 filtros são obtidos para $1 < h < 26$ através do uso de três frequências determinantes para cada filtro: as duas de borda e a frequência central. O filtro $W_h[l]$ é descrito pela equação

$$W_h[l] = \begin{cases} \frac{l - cf_{(h-1)}}{cf_h - cf_{(h-1)}} & cf_{(h-1)} \leq l < cf_h \\ 1 & l = cf_h \\ \frac{cf_{(h+1)} - l}{cf_{(h+1)} - cf_h} & cf_h < l \leq cf_{(h+1)}. \end{cases} \tag{2.8}$$

Logo, a saída do h-ésimo banco de filtro no i-ésimo segmento $fb_h^{(i)}$ é descrita pela equação

$$fb_h^{(i)} = \sum_l W_h[l] F^{(i)}[l], \tag{2.9}$$

em que $W_h[l]$ é a ponderação triangular aplicada a l-ésima frequência do canal do i-ésimo segmento $F^{(i)}[l]$.

Os MFCCs ($\{mfc_z^{(i)}\}_{z=1}^{N_{\text{mfc}}}$) são obtidos para o i-ésimo segmento de fala através do uso da transformada discreta de cosseno (DCT) ao logaritmo das energias do banco de filtro pela equação

$$mfc_z^{(i)} = \sum_{h=1}^{24} \ln fb_h cos(\frac{\pi z(h - 0,5)}{24}), \tag{2.10}$$

em que $z = 1, \ldots, N_{\text{mfc}}$.

2.1.3 Características prosódicas

A prosódia está relacionada com aspectos de entonação e ritmo da linguagem, em particular através do uso de F0, energia e duração para transmitir significados pragmáticos, afetivos e conversacionais (JURAFSKY; MARTIN, 2026). Essas características contêm as informações extraídas da melodia, do ritmo e da ênfase na fala que aumentam a inteligibilidade da mensagem falada, permitindo ao ouvinte segmentar a fala contínua em frases e palavras com facilidade (MARY; YEGNANARAYANA, 2008). Neste trabalho, duas dessas características serão investigadas: a F0 e a energia.

As duas características prosódicas que são extraídas ao longo deste trabalho descrevem informações presente nas modulações dos pulsos de ar que são produzidos pelas pregas vocais no modelo fonte-filtro linear (FANT, 1960). A primeira delas é a F0, que representa a frequência dos pulsos de ar por segundo que passam devido ao movimento de abertura e fechamento da glote. Ela é percebida pelo ouvido através da tonalidade do som, que pode variar do grave ao agudo. A outra característica extraída é a energia, que é percebida pelo ouvido como o volume do som. Quanto maior a pressão subglótica acumulada e mais abrupto o fechamento das pregas vocais, maior será a intensidade acústica dos pulsos produzidos (QUATIERI, 2008). Nos segmentos onde isso ocorre, a energia será mais concentrada. Assim, a combinação de F0 e energia fornece um perfil mais completo dos pulsos em um segmento.

Esse perfil conjunto de F0 e energia fundamenta o desenvolvimento de sistemas para a detecção automática de marcas de diálogo (ANANTHAKRISHNAN; NARAYANAN, 2008). Adicionalmente, esse perfil pode ser mapeado em incorporações vetoriais (“embeddings”), agregando informações cruciais para o desempenho dos modelos de linguagem atuais (PORTEŠ; HORÁK, 2025). Desse modo, mesmo quando utilizadas isoladamente, esse par de características prosódicas exerce um papel central no processamento de fala. Os fundamentos que descrevem o procedimento de extração dessas características são discutidos nas próximas subseções.

2.1.3.1 Energia

A energia acústica da voz é proveniente da energia aerodinâmica subglótica, a qual é responsável pela oscilação das pregas vocais (TITZE, 2001). Desse modo, a produção vocal pode ser compreendida como um processo de conversão energética, no qual a vibração das pregas modula o fluxo de ar glotal que produz o som da voz. Esse som é percebido com maior volume quando a energia acústica da fala é maior (ALLEN, 1971).

A energia do i-ésimo segmento $e^{(i)}$ é proporcional à soma dos quadrados das amostras do segmento (RABINER; SCHAFER, 2011). Portanto, para fins de processamento de sinais, costuma-se definir uma (pseudo) medida de energia adimensional como na equação

$$e^{(i)} = \sum_{n=1}^{M} s^{(i)}[n]^2 w_R[n], \tag{2.11}$$

em que $w_R[n]$ é a janela retangular de tamanho M.

2.1.3.2 F0

A F0 é a frequência da quantidade de pulsos de ar que passam por segundo pela glote devido ao movimento de abertura e fechamento das pregas vocais. Ela é percebida pelo ouvido através da tonalidade do som, que varia do grave ao agudo.

A estimação de F0 através de amostras de áudio pode ser realizada por meio de diversos algoritmos, que geralmente empregam alguma forma de função de correlação (REUTER, 2019). O algoritmo utilizado neste trabalho é o RAPT ("Robust Algorithm for Pitch Tracking") (TALKIN; KLEIJN, 1995). Esse algoritmo foi utilizado em Adami (2007), a principal referência para este trabalho. Assim, ele é adotado para estimar F0.

O algoritmo RAPT é implementado com dois passos centrais (TALKIN; KLEIJN, 1995). O cálculo da função de autocorrelação cruzada normalizada (NCCF) e a seleção de candidatos por programação dinâmica.

O i-ésimo valor da NCCF com atraso τ é representado por $\Phi^{(i)}[\tau]$ que é obtido através da equação

$$\Phi^{(i)}[\tau] = \frac{\sum_{n=1}^{M_1} s^{(i)}[n] \cdot s^{(i)}[n+\tau]}{\sqrt{(\sum_{n=1}^{M_1} s^{(i)}[n]^2) \cdot (\sum_{n=1}^{M_1} s^{(i)}[n+\tau]^2)}}, \tag{2.12}$$

em que M_1 amostras é o tamanho da janela de integração que tem 15 ms de duração. Os valores de τ são calculados entre $\tau_{\min}$ que é associado a 500 Hz e $\tau_{\max}$ que é associado a 50 Hz. Assim, os segmentos tem $M_2 = M_1 + \tau_{\max}$ amostras, mas são obtidos com muita sobreposição a cada 10 ms.

A partir dos valores calculados de $\Phi^{(i)}[\tau]$, o algoritmo RAPT identifica os picos locais que superam um limiar empírico, selecionando os atrasos correspondentes como candidatos a período fundamental para o segmento corrente (TALKIN; KLEIJN, 1995). Adicionalmente, um estado de 'não-vozeado' (silêncio ou consoante surda) é incluído no conjunto de candidatos.

A estimação final do valor de $f0^{(i)}$ não ocorre de forma isolada para cada segmento. Para evitar saltos abruptos e erros de dobra ou divisão por dois na estimação da F0, o RAPT emprega programação dinâmica para encontrar a trajetória de F0 globalmente ideal ao longo de todo o sinal (TALKIN; KLEIJN, 1995). Esse processo consiste em minimizar uma função de custo acumulada que pondera dois fatores:

- **Custo local:** Determinado pela magnitude do pico da NCCF. Quanto maior o valor de $\Phi^{(i)}[\tau]$, menor é o custo de selecionar aquele candidato.

- **Custo de transição:** Penaliza mudanças excessivamente bruscas na frequência fundamental ou transições improváveis entre segmentos vozeados e não-vozeados consecutivos.

Ao final da otimização global, o caminho que minimiza o custo total define o atraso ótimo τ_* para cada segmento (TALKIN; KLEIJN, 1995). Por fim, a estimativa da frequência fundamental para o i-ésimo segmento é obtida pela relação inversa $f0^{(i)} = F_s/\tau_*^{(i)}$, em que F_s é a taxa de amostragem do sinal.

2.1.4 Quantização de características

Após o procedimento de extração de características, o conteúdo de cada segmento de fala é representado por um vetor contínuo. Para mitigar pequenas variações e ruídos intrínsecos a essa representação, é conveniente aplicar uma estratégia de particionamento e discretização do espaço vetorial com base em um modelo de misturas Gaussianas (GMM) (BISHOP; NASRABADI, 2006). Sob essa abordagem probabilística, cada vetor contínuo tem sua verossimilhança avaliada frente às componentes do modelo, sendo mapeado para o agrupamento de maior probabilidade a posteriori (BISHOP; NASRABADI, 2006). Essa operação viabiliza a conversão do conjunto de vetores contínuos de teste em uma sequência discreta de símbolos, utilizando o GMM previamente ajustado como o próprio quantizador do espaço de características.

O GMM é um modelo descrito através da distribuição resultante da soma ponderada de distribuições Gaussianas. Uma distribuição Gaussiana $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ descreve a densidade de probabilidade para um vetor $\mathbf{x}$ com a equação

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}, \tag{2.13}$$

em que $\boldsymbol{\mu}$ é um vetor com a média da Gaussiana, $\boldsymbol{\Sigma}$ é a matriz de covariância e D é a quantidade de dimensões da distribuição (BISHOP; NASRABADI, 2006).

A função de verossimilhança do modelo de mistura de Gaussianas $p(\mathbf{x}; L)$ com K componentes Gaussianas é descrita pela equação

$$p(\mathbf{x}; L) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.14}$$

em que $L = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ é o conjunto formado pelos parâmetros das Gaussianas, que são chamadas de componentes da mistura e π_k é o coeficiente da mistura que pondera o impacto na mistura da componente descrita pela variável k, de tal maneira que $\sum_{k=1}^{K} \pi_k = 1$ (BISHOP; NASRABADI, 2006). Esse modelo pode ser ajustado para descrever distribuições complexas, à medida que quase qualquer densidade contínua pode ser aproximada com uma mistura de Gaussianas.

O ajuste dos parâmetros do GMM pode ser realizado com o algoritmo de expectativa e maximização, o EM. O algoritmo EM refina os parâmetros GMM para aumentar a verossimilhança

do modelo monotonicamente a cada iteração (REYNOLDS; QUATIERI; DUNN, 2000). Uma iteração desse algoritmo é divida em duas etapas, a de expectativa e a de maximização.

Na etapa da expectativa (BISHOP; NASRABADI, 2006), as probabilidades a posteriori $\gamma_k(\mathbf{x}_i)$ são calculadas para indicar a responsabilidade da Gaussiana indicada pelo índice k sobre o vetor de características $\mathbf{x}_i$. O vetor de características $\mathbf{x}_i$ é um dos elementos do conjunto de dados de treinamento $\mathbf{X}_{\text{treino}} = \{\mathbf{x}_i\}_{i=1}^{N}$. Esses dados de treinamento são usados para calcular as probabilidades a posteriori por meio da equação

$$\gamma_k(\mathbf{x}_i) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \tag{2.15}$$

as quais são utilizadas posteriormente na etapa de maximização.

Na etapa de maximização (BISHOP; NASRABADI, 2006), os parâmetros da mistura de Gaussianas são atualizados através das seguintes equações:

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \gamma_k(\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{N} \gamma_k(\mathbf{x}_i)}, \tag{2.16}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} \gamma_k(\mathbf{x}_i)(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} \gamma_k(\mathbf{x}_i)}, \tag{2.17}$$

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \gamma_k(\mathbf{x}_i), \tag{2.18}$$

em que k é o índice de uma das K componentes da mistura. Essa atualização dos parâmetros na última etapa finaliza uma iteração do algoritmo EM. Então, as etapas de expectativa e maximização são executadas alternadamente até a convergência do algoritmo, garantindo o aumento local da verossimilhança dos dados.

O ajuste do modelo está completo após a sua convergência no conjunto de treinamento $\mathbf{X}_{\text{treino}}$. Após essa etapa de treinamento, qualquer amostra de teste $\mathbf{x}$ pode ser convertida em um símbolo discreto $\hat{k}$ (i.e., quantizada) correspondente à componente Gaussiana com maior probabilidade *a posteriori*, conforme descrito pela equação

$$\hat{k} = \arg\max_k \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.19}$$

Neste trabalho, esse procedimento é usado para quantizar, de maneira independente, os vetores de MFCCs e de características prosódicas, convertendo-os em duas variáveis aleatórias discretas.

2.2 Sobre entropia e o problema de estimação

Após o processo de quantização, o conjunto de MFCCs e as característica prosódicas são ambos codificados como variáveis aleatórias discretas, que denotaremos por X e Y, respectivamente.

O ponto central deste trabalho é investigar a hipótese de que essas duas variáveis compartilham algum tipo de dependência estatística. Nesse sentido, o grau de dependência estatística entre duas variáveis X e Y pode ser quantificado de diversas formas.

Em particular, para este trabalho, foi feita a opção pelo uso da medida de entropia condicional $H(Y|X)$, que fornece a quantidade de média da incerteza restante acerca do valor de Y uma vez que X é conhecido, como enunciado por Shannon (1948).

Dessa forma, esta seção revisita conceitos de teoria da informação, trazendo as definições pertinentes para este trabalho. Além disso, é introduzido o problema de estimação de entropia (condicional) dado um conjunto de instâncias independentes das variáveis de interesse.

2.2.1 Entropia e Entropia Condicional

Dada uma variável aleatória X, que pode assumir qualquer valor no espaço amostral $\Omega_x = \{1, \ldots, K_x\}$, com distribuição $P = (p_1, \ldots, p_{K_x})$, sua entropia de Shannon é dada por

$$H(X) = -\sum_{i=1}^{K_x} p_i \log(p_i). \tag{2.20}$$

Essa medida fornece uma espécie de quantificação do grau de incerteza acerca do valor de X antes da realização do experimento aleatório. Quando medida em bits de informação (i.e. quando a base do logaritmo é 2), $H(X)$ informa que prever o valor de X é tão difícil quanto prever o valor de um experimento com C resultados possíveis e uniformemente distribuídos, em que C é a cardinalidade efetiva de X, conforme discutido em Montalvao, Canuto e Carvalho (2016), descrita por

$$C = 2^{H(X)}. \tag{2.21}$$

Analogamente, para definir a entropia condicionada, considere também uma segunda variável Y com espaço amostral $\Omega_y = \{1, \ldots, K_y\}$ e distribuição $Q = (q_1, \ldots, q_{K_y})$. A entropia condicional de X dado Y é dada por

$$H(X|Y) = -\sum_{j=1}^{K_y} q_j \sum_{i=1}^{K_x} p_{ij} \log p_{ij}, \tag{2.22}$$

em que $p_{ij} = Pr(X = i|Y = j)$.

Enquanto indicador de dependência, $H(X|Y)$ satisfaz as seguintes propriedades:

1. $H(X|Y) \leq H(X)$, com a igualdade sendo satisfeita se e somente se X e Y forem independentes.

2. $H(X|Y) = 0$ quando X for funcionalmente dependente de Y. Isto é, quando o conhecimento de Y informa o valor de X sem ambiguidades.

3. Valores intermediários no intervalo $[0, H(X)]$ indicam níveis intermediários de dependência.

Assim como no caso não condicionado, pode-se recorrer a noção de cardinalidade efetiva para a entropia condicionada (Equação 2.21). Especificamente, $C_{X|Y} = 2^{H(X|Y)}$ (quando a entropia é medida em bits) informa que prever o resultado X dado o resultado de Y é, em média, tão difícil quanto prever um resultado de um experimento com $C_{X|Y}$ símbolos equiprováveis.

2.2.2 Estimação de entropia

Determinar a entropia de uma variável aleatória X requer o conhecimento de sua distribuição. Entretanto, em situações realistas, as verdadeiras probabilidades quase nunca são conhecidas. Dessa forma, faz-se necessário adotar um estimador $\hat{H}$, que recebe como entrada um conjunto de N instâncias independentes de X, denotado por $\{x^{(i)}\}_{i=1}^{N}$, e retorna como saída um escalar que busca aproximar a verdadeira entropia (i.e. $\hat{H}(X) \approx H(X)$).

A forma mais natural de estimar a entropia $H(X)$ é estimando empiricamente as probabilidades de cada elemento do espaço amostral ($\hat{p}_i = N_i/N$, em que N_i é a quantidade de vezes que o resultado i foi observado), e substituí-las diretamente na Equação 2.20.

Infelizmente, essa abordagem, conhecida como *plug in*, costuma resultar em estimativas altamente enviesadas, quando o tamanho da amostra N não é grande o suficiente, como bem documentado na literatura (MILLER, 1955). Dessa forma, adotar estimadores mais precisos é uma preocupação relevante em estudos nos quais a quantidade de dados é limitada. Para este estudo, em particular, adota-se o estimador Chao-Shen (CHAO; SHEN, 2003), que foi consistentemente classificado entre os estimadores mais precisos em experimentos com tamanhos de amostra e espaço amostral variados, realizados em um estudo comparativo em (TORRE et al., 2025).

2.2.3 O estimador Chao-Shen

O estimador proposto por Chao e Shen combina duas ideias fundamentais. A primeira delas consiste na aplicação do estimador de Horvitz-Thompson para estimação de entropia, que tem como atributo mais atraente o fato de que ele considera os símbolos não observados na amostra.

Considere que os símbolos $\{1, \ldots, K_x\}$ estão associados à quantidades de interesse $\mu_1, \ldots, \mu_{K_x}$, respectivamente. Suponha uma amostra com N instâncias em que k símbolos diferentes foram observados, que serão denotados por $j_1, \ldots, j_k$, em que $k \leq K_x$. Para cada símbolo i, denotamos por N_i a sua frequência observada na amostra, de modo que $\sum_i N_i = N$.

Um estimador não enviesado para $\tau = \sum_{i=1}^{K_x} \mu_i$, introduzido por Horvitz e Thompson (1952), é dado por

$$\hat{\tau} = \sum_{i=1}^{k} \frac{\mu_{j_i}}{\theta_{j_i}}, \tag{2.23}$$

em que θ_m é a probabilidade de que o símbolo m esteja incluído em uma amostra qualquer com N instâncias. Se considerarmos $\mu_i = -p_i \log(p_i)$, uma estimativa para a entropia pode ser obtida usando o estimador de Horvitz-Thompson:

$$\hat{H} = \sum_{i=1}^{k} \frac{-p_i \log(p_i)}{1 - (1 - p_i)^N}. \tag{2.24}$$

O segundo elemento fundamental do estimador Chao-Shen são estimativas de probabilidade corrigidas usando o conceito de cobertura amostral, definido por

$$A = 1 - \sum_{i \; t.q \; N_i = 0} p_i. \tag{2.25}$$

Em palavras, A é a massa de probabilidade total dos símbolos que foram observados pelo menos uma vez na amostra. Analogamente, $1 - A$ é a massa de probabilidade faltante, isto é, a probabilidade de que se mais uma instância for coletada , esta será de um símbolo ainda não observado dentre as N instâncias anteriores. Um estimador muito conhecido para A foi introduzido por (GOOD, 1953), dado por

$$\hat{A} = 1 - \frac{h_1}{N}, \tag{2.26}$$

em que h_1 é a quantidade de símbolos "solitários", isto é, que apareceram uma única vez na amostra ($N_i = 1$). Esse é um resultado bastante intuitivo, pois a ocorrência de um símbolo não observado na próxima instância equivale à adição de mais um símbolo "solitário". Em (CHAO; SHEN, 2003), os autores consideram que a estimativa empírica N_i/N é uma boa estimativa para $Pr(X = i | X \in \{j_1, \ldots, j_k\})$, e não para $Pr(X = i)$. Pela regra de Bayes, segue que

$$Pr(X = i | X \in \{j_1, \ldots, j_k\}) = \frac{Pr(X \in \{j_1, \ldots, j_k\} | X = i) \cdot Pr(X = i)}{Pr(X \in \{j_1, \ldots, j_k\})}. \tag{2.27}$$

Note que $Pr(X \in \{j_1, \ldots, j_k\} | X = i) = 1$, uma vez que o interesse é em estimar a probabilidade dos símbolos que foram observados na amostra, e que $Pr(X \in \{j_1, \ldots, j_k\}) = A$. Logo, p_i pode ser estimado por

$$\hat{p}_i = \frac{N_i}{N} \cdot \hat{A}. \tag{2.28}$$

Finalmente, o estimador Chao-Shen substitui as probabilidades estimadas na Equação 2.28 no estimador de entropia da Equação 2.24. Ao longo deste trabalho, esse estimador é utilizado para estimar as entropias dos MFCCs (quantizados) condicionados à prosódia (quantizada) – ou vice-versa.

3 Métodos e Materiais

Este capítulo detalha como os testes de hipóteses deste trabalho foram realizados. A primeira seção descreve a definição dos testes de hipótese nula, enquanto a segunda seção detalha o protocolo experimental que foi utilizado para implementá-los. Por fim, a terceira seção destaca limitações no escopo do trabalho.

3.1 Teste de Hipótese Nula

Como mencionado anteriormente, o objetivo deste trabalho é avaliar a hipótese de independência entre os MFCCs – codificados pela variável aleatória X – e as características prosódicas – codificadas pela variável aleatória Y – de um trecho de fala. Mais especificamente, busca-se averiguar se a ordenação natural dos pares de amostras de X e Y impacta a dependência entre eles. A suposição de independência estatística entre as variáveis em sua ordem natural constitui a hipótese nula H_0. Dada uma sequência de observações de X, denotada por $S_X = \{x^{(i)}\}_{i=1}^{N}$, e uma sequência pareada de observações de Y, denotada por $S_{Y_{\text{test}}} = \{y_{\text{test}}^{(i)}\}_{i=1}^{N}$, em que $x^{(i)}$ e $y^{(i)}$ codificam as observações correspondentes ao i-ésimo segmento de fala em um conjunto de dados, o grau de dependência entre X e Y pode ser quantificado de diversas maneiras.

Para este propósito, adotou-se a medida de entropia condicional $H(Y|X)$, que indica a quantidade média de incerteza sobre Y que permanece após o conhecimento da variável X, conforme explicado na Seção 2.2.1. Espera-se que, quando X e Y são dependentes, o valor de $H(Y|X)$ seja menor do que quando as variáveis são independentes. Essa medida, quando expressa em bits (de informação), pode ainda ser associada à noção de cardinalidade efetiva descrita por

$$C_{Y|X} = 2^{H(Y|X)}. \tag{3.1}$$

Uma avaliação da significância estatística de $C_{Y|X,\text{test}}$ – a cardinalidade efetiva estimada utilizando as sequências S_X e $S_{Y_{\text{test}}}$ – indicará se as evidências contrariam ou não a hipótese nula (BIAU; JOLLES; PORCHER, 2010). Essa avaliação realiza-se mediante a estimativa empírica da distribuição $F(c) = \Pr(C_{Y|X} < c|H_0)$, a qual pressupõe a validade de H_0.

Para estimar a distribuição condicional $F(c)$, é necessário obter um conjunto de cardinalidades efetivas estimadas a partir de diferentes sequências de amostras de X e Y em que H_0 é verdadeira. Uma estratégia simples para obter tais sequências consiste no embaralhamento aleatório da sequência de teste $S_{Y_{\text{test}}}$. Cada permutação produzirá uma sequência distinta de características prosódicas S_Y. Assim, cada $x^{(i)}$ será associado a uma característica prosódica aleatória de outro segmento de áudio, garantindo a independência entre X e Y. Portanto, $F(c)$ pode ser calculada por meio de estimativas de $C_{Y|X}$ com S_X constante e diferentes permutações

de S_Y. Os detalhes do procedimento utilizando o estimador Chao-Shen para a obtenção das amostras de $C_{Y|X}$ são apresentados a seguir.

O primeiro passo consiste em estimar a probabilidade empírica de cada valor discreto possível de X por meio da sequência $S_X = \{x^{(1)}, \ldots, x^{(N)}\}$, em que $x^{(i)} \in A$. Aqui, $A = \{a_1, a_2, \ldots, a_G\}$ representa o conjunto de valores discretos que X pode assumir. Então, a probabilidade que X assuma o valor a_j é estimada por

$$\hat{\Pr}(X = a_j) = \frac{\sum_{i=1}^{N} I_{x^{(i)}=a_j}}{N}, \tag{3.2}$$

em que $I_{x^{(i)}=a_j}$ é definida por

$$I_{x^{(i)}=a_j} = \begin{cases} 1, \text{ se } x^{(i)} = a_j \\ 0, \text{ caso contrário} \end{cases}. \tag{3.3}$$

O próximo passo utiliza uma permutação aleatória dos N elementos de $S_{Y_{\text{test}}}$ para gerar S_Y, uma sequência dada por $S_Y = \{y^{(1)}, y^{(2)}, ..., y^{(N)}\}$, em que $y^{(i)} \in B$. O conjunto $B = \{b_1, b_2, ..., b_L\}$ delimita os valores discretos passíveis de serem assumidos por Y. De forma análoga, a probabilidade condicional $\Pr(Y = b_l|X = a_j)$ – também denotada por $\Pr(b_l|X = a_j)$ – de Y assumir o valor b_l quando X assume o valor a_j é estimada por

$$\hat{\Pr}(b_l|X = a_j) = \frac{\sum_{i=1}^{N} I_{y^{(i)}=b_l} \cdot I_{x^{(i)}=a_j}}{\sum_{i=1}^{N} I_{x^{(i)}=a_j}}. \tag{3.4}$$

Na etapa subsequente, obtêm-se as estimativas das entropias condicionais marginais $H(Y|X = a_j)$, para todo $j = 1, \ldots, G$. Note que os tamanhos das amostras utilizados para estimar cada uma dessas entropias condicionais correspondem à frequência com que cada evento da forma $X = a_j$ é observado no conjunto de dados. Essa quantidade pode ser insuficiente para uma estimativa empírica adequada da entropia devido às limitações do estimador *plug in*. Por esse motivo, optou-se pelo estimador Chao-Shen, dada a sua eficiência demonstrada em cenários de baixa amostragem, conforme destacado na Seção 2.2.3. As estimativas das entropias condicionais sob o estimador de Chao-Shen são dadas por

$$\hat{H}(Y|X = a_j) = -\sum_{l=1}^{L} \frac{\hat{\Pr}_{gt}(b_l|X = a_j) \log_2 \hat{\Pr}_{gt}(b_l|X = a_j)}{1 - (1 - \hat{\Pr}_{gt}(b_l|X = a_j))^{n_j}}, \tag{3.5}$$

em que n_j é o número de observações do evento $X = a_j$ e $\hat{\Pr}_{gt}(b_l|X = a_j)$ representam as estimativas de probabilidade corrigidas pelo método de Good-Turing, dadas por

$$\hat{\Pr}_{gt}(b_l|X = a_j) = \left(1 - \frac{m_j}{n_j}\right) \hat{\Pr}(b_l|X = a_j), \tag{3.6}$$

em que m_j é o número de símbolos solitários na amostra, isto é, a quantidade de eventos da forma $Y = b_l$ observados apenas uma vez quando $X = a_j$.

A média de todas as entropias condicionais marginais de Y resulta na entropia condicional $H(Y|X)$, estimada como

$$\hat{H}(Y|X) = \sum_{j=1}^{M} \hat{\Pr}(X = a_j)\hat{H}(Y|X = a_j). \tag{3.7}$$

Essa medida é finalmente empregada para calcular a cardinalidade efetiva $C_{Y|X}$ por meio da equação

$$\hat{C}_{Y|X} = 2^{\hat{H}(Y|X)}. \tag{3.8}$$

Assim, ao repetir o procedimento acima para D permutações diferentes de $S_{Y_{\text{test}}}$, obtém-se uma amostra de cardinalidades efetivas $C^N_{Y|X} = \{\hat{C_{Y|X}}^{(1)}, ..., \hat{C_{Y|X}}^{(d)}, ..., \hat{C_{Y|X}}^{(D)}\}$ na qual sabe-se que a hipótese H_0 se sustenta. Por meio do estimador Chao-Shen, também se pode obter a cardinalidade efetiva de teste $\hat{C}_{Y|X,\text{test}}$, com a sequência de teste.

Finalmente, obtém-se uma estimativa da probabilidade de observar um resultado tão extremo quanto, ou mais extremo do que, aquele observado, sob a hipótese nula, através da equação

$$\hat{Pr}(C_{Y|X} \leq \hat{C}_{Y|X,\text{test}}|H_0) = \frac{\sum_{d=1}^{D} I_{\hat{C_{Y|X}}^{(d)} \leq C_{Y|X,\text{test}}}}{D}. \tag{3.9}$$

Essa probabilidade é denominada valor-p do teste (LEHMANN; ROMANO, 2005).

3.2 Protocolo Experimental

O protocolo experimental descrito a seguir divide-se em cinco etapas. Inicialmente, apresenta-se o conjunto de dados utilizado na implementação dos testes. Em seguida, detalha-se o procedimento de extração de características dos dados. Na terceira etapa, os vetores resultantes são convertidos em símbolos discretos via quantização vetorial. A quarta etapa aborda a organização desses símbolos para a execução dos testes. Por fim, descreve-se o procedimento para realizar os testes propostos neste estudo.

3.2.1 Conjunto de Dados

O presente estudo utilizou o conjunto de dados que foi introduzido em Ynoguti (1999), composto por 1600 arquivos de áudio contendo gravações de 40 oradores. Cada orador pronunciou um conjunto de 40 frases em português brasileiro. Essas frases foram selecionadas de um repertório de 200 sentenças foneticamente balanceadas (ALCAIM; SOLEWICZ; MORAES, 1992). Desse modo, cada um dos 5 conjuntos de frases foi falado por 4 oradores diferentes, 2 homens e 2 mulheres. Portanto, o conjunto de dados apresenta equilíbrio de gênero, contando com igual número de oradores do sexo masculino e feminino.

A escolha específica deste conjunto justifica-se pela sua natureza estritamente controlada, o que facilita a extração de conclusões em uma investigação em estágio inicial. Além do controle

linguístico e de gênero, as gravações foram realizadas em ambiente silencioso, com o uso de um microfone direcional (YNOGUTI, 1999). Esse controle de ambiente e equipamento garante que a correlação entre as características não seja corrompida ou induzida por fatores externos, isolando as variáveis de interesse do experimento.

As gravações possuem a frequência de amostragem F_s de 11,025 kHz. Cada gravação possui, em média, 3 segundos de duração, totalizando aproximadamente 75 minutos de sinal de fala no conjunto.

3.2.2 Extração de Características

A extração de características implementada neste trabalho reproduz o procedimento proposto em Adami (2007). Nele, todo sinal de fala $s[n]$ do conjunto utilizado foi processado para ser representado por dois conjuntos de vetores: vetores de MFCCs e vetores de prosódia. Esse último conjunto foi obtido a partir dos valores de energia e F0, extraídos por meio de um processo de janelamento com 10 ms duração e sem sobreposição.

Primeiramente, as estimativas de F0 foram obtidas com uma implementação do algoritmo RAPT fornecida pelo pacote "VOICEBOX" (BROOKES, 1997), visto que os detalhes operacionais desse algoritmo extrapolam o escopo deste trabalho. Essa implementação do RAPT forneceu, para cada segmento janelado, uma estimativa do valor de F0 ou um indicador de não vozeamento. Todos os segmentos com F0 também tiveram sua energia $e^{(i)}$ extraída. Logo, nesta primeira etapa, o i-ésimo segmento com fala vozeada foi representado pelo o vetor de características prosódicas $(e^{(i)}, f0^{(i)})$.

O passo seguinte consistiu no cálculo da tendência de variação no i-ésimo segmento, gerando o vetor $(\Delta e^{(i)}, \Delta f0^{(i)})$. Esse cálculo foi realizado por meio da equação

$$(\Delta e^{(i)}, \Delta f0^{(i)}) = \frac{\sum_{\tilde{\imath}=i-I}^{i+I} \tilde{\imath}\{e^{(i)}, f0^{(i)}\}}{\sum_{\tilde{\imath}=i-I}^{i+I} \tilde{\imath}^2}, \tag{3.10}$$

em que $I = 2$. Assim, os segmentos vozeados inseridos em trechos com menos de 5 segmentos contínuos foram descartados da análise, da mesma forma que os segmentos não vozeados. Os segmentos remanescentes foram descritos por um vetor de características prosódicas com 4 dimensões, após a concatenação dos valores brutos e da variação deles: $(e^{(i)}, f0^{(i)}, \Delta e^{(i)}, \Delta f0^{(i)})$.

Seguindo a metodologia de Adami (2007), a segunda etapa consistiu na extração de 19 MFCCs de cada segmento preservado (i.e., segmentos vozeados em trechos contendo mais de 4 segmentos vozeados), com uma largura de banda definida por $cf_1 = 300$ Hz e $cf_{26} = 3138$ Hz. Os trechos foram segmentados utilizando janelas de Hamming de 20 ms de duração com 50% de sobreposição. Dessa forma, os segmentos alinharam-se temporalmente aos segmentos das características prosódicas, uma vez que ambos possuem um intervalos de deslocamento de 10 ms. O processo de extração desses coeficientes seguiu a explicação presente na Seção 2.1.2.

Além dos MFCCs, extraiu-se a variação temporal desses coeficientes, que foi posteriormente concatenada ao vetor de coeficientes brutos. Com isso, obteve-se um conjunto de segmentos em que cada unidade está associada a um par de vetores: um vetor de características prosódicas (e suas variações) pertencente ao $\mathbb{R}^4$ e um vetor de MFCCs (e suas variações) pertencente ao $\mathbb{R}^{38}$. Doravante, ao longo deste capítulo, o termo "segmento" designará o par de vetores correspondente.

3.2.3 Quantização de Características

Como resultado do processo descrito na subseção anterior, os vetores de características obtidos apresentam natureza contínua. Para modelar a distribuição desses vetores, o estado da arte da literatura de reconhecimento de orador já recorreu ao uso de GMM (REYNOLDS; QUATIERI; DUNN, 2000; SHARMA et al., 2024). No presente trabalho, a escolha desse modelo justificou-se por ele fornecer uma quantização vetorial que viabilizou a investigação da dependência entre símbolos discretos. Assim, os vetores foram quantizados por meio de GMMs, com parâmetros ajustados pelo algoritmo EM, conforme explicado na Seção 2.1.4. Porém, o ajuste desses modelos esteve sujeito a restrições particulares.

Especificamente, a quantização dos segmentos associados as amostras de teste foi feita a partir de um modelo ajustado para um outro conjunto de oradores. Caso contrário, o vazamento da informação de identidade dos oradores criaria uma dependência artificial entre os vetores de MFCCs e os de prosódia.

Para lidar com essa restrição, os 40 oradores da base foram particionados aleatoriamente de maneira uniforme em dois grupos com balanceamento de gênero. Desse modo, 20 oradores foram selecionados para o grupo de teste e o restante foi separado em um grupo de treinamento. Portanto, os segmentos das falas deste último foram separados em um conjunto de treinamento X_{treino} — usado para ajustar os GMMs —, enquanto o restante dos segmentos foi separado no conjunto X_{teste} — usado durante os testes de hipótese.

Assim, utilizando o EM, dois GMMs foram ajustados a partir das amostras de cada segmento do conjunto X_{treino}. O primeiro modelou os vetores de MFCCs (e suas variações), por meio de uma mistura de $K_{\text{MFCC}} = 2048$ componentes Gaussianas. O segundo, por sua vez, modelou os vetores prosódicos, com $K_{\text{pros}} = 512$ componentes. Essas quantidades de componentes foram escolhidas para realizar uma validação empírica baseada no artigo de Adami (2007), que utilizou essa mesma configuração e motivou a investigação deste trabalho sobre a independência das características.

Após o ajuste, os pares de vetores do conjunto X_{teste} foram quantizados por meio da regra do máximo a posteriori, conforme detalhado na Seção 2.1.4. Assim, cada par de vetores nesse conjunto foi mapeado em um par de números inteiros (x, y) pertencente ao produto cartesiano $\{1, \ldots, 2048\} \times \{1, \ldots, 512\}$. A distorção desse mapeamento foi quantificada por meio da

relação sinal-ruído de quantização (SNR_Q), expressa em decibéis (dB).

A SNR_Q avalia distorção introduzida ao mapear a representação vetorial contínua em um conjunto discreto de índices associados às componentes Gaussianas do GMM (PROAKIS; MANOLAKIS, 2013). Ela foi obtida através do cálculo do erro vetorial er, definido pela equação

$$er = v - \mu_{\hat{c}}, \tag{3.11}$$

que representa o ruído de quantização provocado por substituir o vetor contínuo v pela média do componente com maior verossimilhança $\mu_{\hat{c}}$. Dessa forma, a SNR_Q foi calculada através da equação

$$SNR_Q = 10\log_{10}(E_{sinal}/E_{ruido}) \tag{3.12}$$

em que $E_{sinal} = \sum_{d=1}^{D}(v_d)^2$ é a energia da representação vetorial contínua, $E_{ruido} = \sum_{d=1}^{D}(er_d)^2$ é a energia do ruído de quantização e D é a quantidade de dimensões da representação vetorial. A média dos valores calculados de SNR_Q foi usada para avaliar a qualidade do processo de quantização das características.

O valor médio da SNR_Q depende diretamente da dimensão efetivamente ocupada pelo sinal. Por exemplo, para sinais distribuídos uniformemente em uma dimensão (mesmo que sua dimensão aparente seja maior que 1), espera-se, em média, que K centróides de quantização distribuídos sobre essa dimensão intrínseca (DI) unitária segmentem o espaço em K intervalos ($K + 1$, a rigor, mas essa unidade pode ser desprezada por simplicidade). Esse processo produz erros de quantização não maiores que metade desses intervalos, logo, da ordem de $\frac{1}{2K}$ do desvio padrão do sinal não quantizado. Assim, a SNR_Q média será aproximadamente dada por $10\log_{10}((2K)^2) \approx 6 + 20\log_{10}(K)$. Por outro lado, se a DI do sinal for igual a 2, os mesmos K centróides produzirão, em média, $K^{(1/2)}$ intervalos em cada dimensão, e os erros de quantização esperados serão da ordem de $\frac{1}{2K^{(1/2)}}$. Em geral, para uma dada DI, os erros de quantização esperados serão da ordem de $\frac{1}{2K^{(1/DI)}}$, logo, o valor médio da SNR_Q será aproximadamente dado por $10\log_{10}((2K^{(1/DI)})^2) \approx 6 + \frac{20}{DI}\log_{10}(K)$.

3.2.4 Preparação das Sequências

Após o processo de quantização, as falas dos oradores selecionados para os testes ficaram descritas por duas sequências de símbolos. Essas sequências, associadas às 40 falas de um mesmo orador, foram concatenadas para obter duas sequências finais de cada indivíduo: uma sequência S_X, com os índices associados a quantização de vetores de MFCCs, e uma sequência $S_{Y_{\text{teste}}}$, decorrente da quantização dos vetores prosódicos. A partir dessa última sequência, $D = 10000$ versões da sequência S_Y foram geradas por meio de D permutações aleatórias de $S_{Y_{\text{teste}}}$. Desse modo, diferentes sequências S_Y foram obtidas de maneira independente da sequência S_X.

3.2.5 Implementação dos Testes

No total, foram realizados 20 testes — um para cada orador sorteado para o grupo de teste. Todos os testes foram executados seguindo as cinco etapas descritas abaixo:

(i). **Seleção do orador:** Escolha do orador pertencente ao grupo de teste.

(ii). **Extração e quantização de características:** Execução dos procedimentos descritos nas subseções 3.2.2 e 3.2.3, respectivamente, para a obtenção das sequências S_X e $S_{Y_{\text{teste}}}$, conforme definidas na Seção 3.1.

(iii). **Estimação das probabilidades empíricas:** Cálculo das probabilidades $\hat{\Pr}(X = a_j)$ a partir da sequência S_X, contendo os valores quantizados dos MFCCs do orador selecionado, utilizando-se a Equação 3.2.

(iv). **Obtenção das cardinalidades efetivas:** Determinação das cardinalidades efetivas sob a hipótese nula a partir das $D = 10^4$ permutações diferentes de $S_{Y_{\text{teste}}}$, utilizando-se a Equação 3.8. Nesta etapa, a cardinalidade efetiva original $\hat{C}_{Y|X,\text{teste}}$ também foi obtida a partir da sequência $S_{Y_{\text{teste}}}$.

(v). **Cálculo do limite superior para o valor-p:** Obtenção do valor-p limite por meio da aplicação da Equação 3.9.

3.3 Limitações do estudo

Apesar do rigor metodológico adotado na formulação dos testes de hipóteses e na extração das características, os resultados obtidos neste trabalho devem ser interpretados à luz de certas limitações inerentes ao protocolo experimental e aos dados utilizados.

Primeiramente, o número de componentes Gaussianas utilizado na quantização foi predefinido com base nos valores adotados por Adami (2007) (i.e., $K_{\text{MFCC}} = 2048$ e $K_{\text{pros}} = 512$). Essa quantidade foi mantida fixa para viabilizar o uso dos mesmos dados testados na primeira parte durante a reprodução do experimento biométrico da segunda parte. No entanto, a quantização pode apresentar algum impacto na dependência entre os dois conjuntos de características. A investigação sobre esse impacto ficou para futuros trabalhos, uma vez que o escopo do atual trabalho se caracteriza como uma prova de conceito.

O escopo deste trabalho também excluiu de sua análise o impacto do ruído ambiente na relação de dependência entre as características. O ruído ambiente é um fenômeno espectral que seria capturado diretamente pelos MFCCs. Além disso, a sua presença provocaria uma mudança na prosódia conhecida como Efeito Lombard (BRUMM; ZOLLINGER, 2011). Isso ocorre porque, em situações ruidosas, os falantes tendem a alterar o esforço vocal, modificando

a energia e F0 de sua fala para melhorar a inteligibilidade. Dessa forma, a presença de ruído modificaria a relação de dependência entre os MFCCs e as características prosódicas.

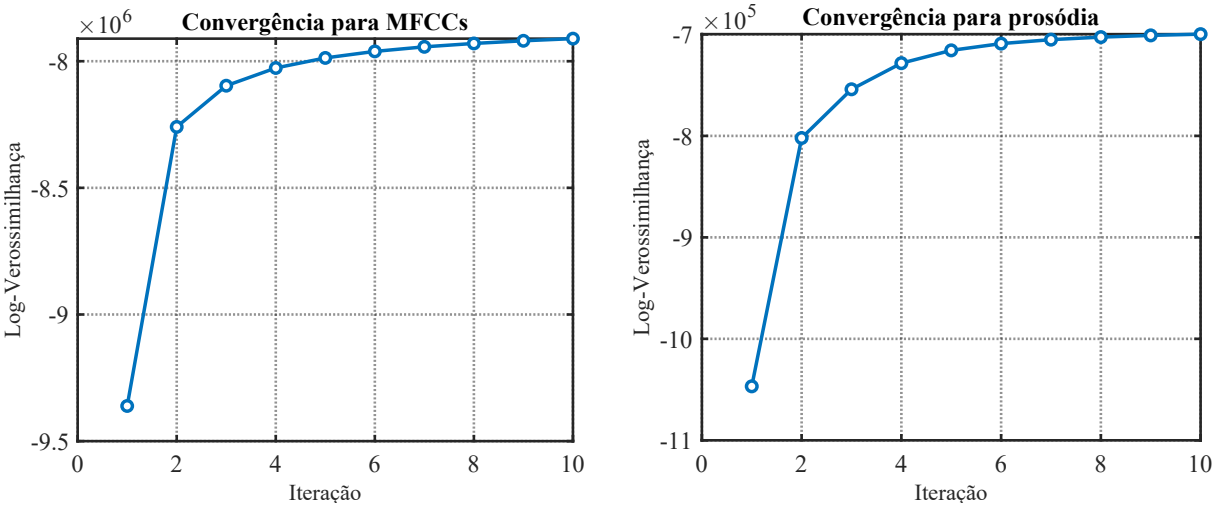
Por fim, a metodologia proposta — baseada em entropia condicional e testes de permutação — é capaz de estimar o grau de dependência entre os conjuntos e a sua significância estatística, porém ela não elucida os mecanismos causais dessa relação. Assim, os elementos fisiológicos ou linguísticos responsáveis pela dependência não ficam explicitados após a realização dos testes. Por conseguinte, a dependência identificada pode ser decorrente de diferentes fatores, exigindo cautela na extrapolação dos resultados. Esta extrapolação pressupõe que as falas sejam obtidas sob as mesmas condições do conjunto testado; generalizações para falas espontâneas, cenários ruidosos ou outros idiomas requerem novos ensaios experimentais.

4 Resultados, Experimento e Discussão

Este capítulo apresenta, em sua seção inicial, os resultados dos testes de hipótese nula conduzidos segundo a metodologia previamente descrita. Nessa seção, modelos foram treinados para quantizar os vetores de prosódia e de MFCCs, viabilizando a avaliação da significância estatística da dependência entre os dois conjuntos de características. Na sequência, o impacto prático dessa dependência é verificado por meio de um experimento de reconhecimento de orador. Esse experimento foi implementado com cada característica e os valores obtidos nas implementações são fundidos com uma rede neural. O resultado dessa fusão reforça a existência da dependência entre as características.

4.1 Resultados para testes de hipótese

Para analisar estatisticamente a dependência entre as características e atender ao objetivo específico de modelar os espaços de características de fala, foi implementada a quantização vetorial dos vetores prosódicos e dos MFCCs utilizando GMMs treinados via EM. Neste trabalho, os modelos convergiram no treinamento com aproximadamente 5 iterações, conforme ilustrado na Figura 1. Esses GMMs foram utilizados para converter os vetores dos oradores testados em símbolos. Contudo, para garantir a qualidade dessa conversão, foi necessário avaliar se esses modelos não apresentavam vieses em relação a nenhum dos indivíduos testados.



(a) Log-Verossimilhança para MFCCs. (b) Log-Verossimilhança para prosódia.

Figura 1 – Comparação dos resultados de convergência do algoritmo EM.

A avaliação dos modelos foi realizada por meio da análise da média e do desvio-padrão da log-verossimilhança para cada orador testado. Como ilustrado na Figura 2, o modelo com 2048 componentes apresentou médias individuais próximas à média global, além de intervalos

de dispersão que se sobrepõem entre os indivíduos. A ausência de discrepâncias significativas indica que o treinamento foi bem-sucedido resultando em uma quantização no espaço de MFCCs independente dos oradores.

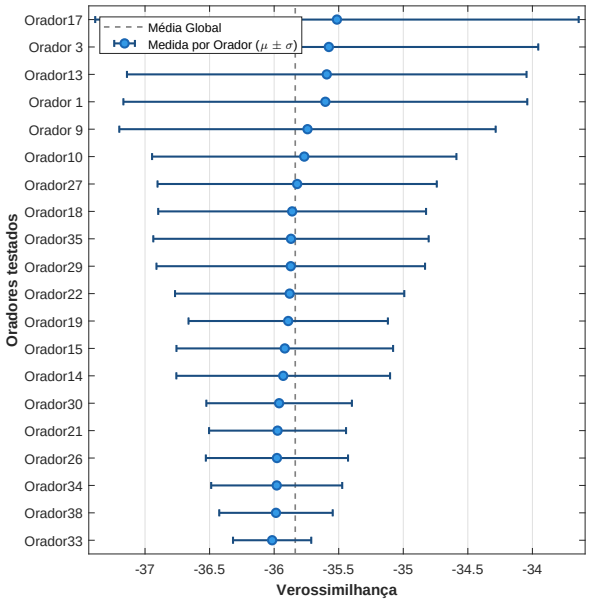


Figura 2 – Média e desvio padrão de Log-Verossimilhança de cada orador para modelo com 2048 componentes.

A independência do modelo com 512 componentes (referente aos vetores prosódicos) foi avaliada sob o mesmo critério. Conforme ilustrado na Figura 3, o modelo apresentou comportamento estatístico equivalente ao observado na Figura 2. Esses resultados confirmaram que ambos os modelos não estavam enviesados. Assim, eles foram utilizados para converter os vetores dos oradores de teste em símbolos.

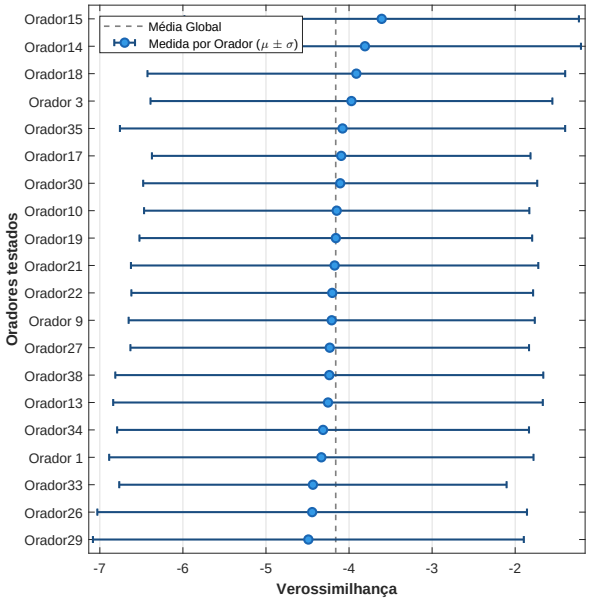


Figura 3 – Média e desvio padrão de Log-Verossimilhança de cada orador para modelo com 512 componentes.

Contudo, antes de realizar os testes de hipóteses, as médias das relações sinal-ruído de quantização SNR_Q foram analisadas. O valor esperado para os MFCCs seria em torno de 7,7 dB se eles estivessem ocupando uniformemente as 38 dimensões aparentes disponíveis. No entanto, o valor obtido experimentalmente com a discretização por 2048 centroides foi de 2,6 dB em média. Ao passo que a média de SNR_Q obtida experimentalmente para a discretização dos vetores prosódicos com 512 centroides foi de 14,3 dB, enquanto a ocupação uniforme das 4 dimensões aparentes produziria uma média de 19,5 dB.

Os resultados empíricos obtidos estão abaixo do esperado nos dois casos, o que sugere que a idealização de distribuições uniformes não é atendida. De fato, no caso de distribuições Gaussianas, por exemplo, o SNR_Q médio esperado para $K = 2048$ e DI = 38 cai para apenas 0,5 dB, em lugar de 7,7 dB. Portanto, o resultado empírico de 2,6 dB indica que a distribuição real dos vetores de MFCCs se situa entre a uniforme e a Gaussiana. O mesmo parece valer para a distribuição do sinal de vetores de prosódia, posto que a média de SNR_Q esperada para $K = 512$ e DI = 4, no caso de distribuição Gaussiana, cai para 12,3 dB, em lugar de 19,5 dB.

O SNR_Q médio empírico de 14,3 dB na discretização dos vetores prosódicos indica ainda que a energia do erro de quantização cometido é aproximadamente 27 vezes menor que a energia do sinal original. Em contraste, para os vetores de MFCC, o SNR_Q médio empírico de apenas 2,6 dB indica que a energia do erro de quantização cometido é de aproximadamente 0,55 da energia do sinal original, o que poderia sugerir a necessidade de um refinamento na quantização desse sinal. No entanto, 2048 centróides são considerados uma escolha empiricamente satisfatória na literatura voltada à biometria da fala. Assim, dois esquemas de quantização podem ser considerados satisfatórios para a representação adequada dos sinais de interesse. Porém, antes de realizar os testes de hipóteses convém analisar a quantidade de informação contida nos sinais resultantes da quantização vetorial.

Primeiramente, a informação da discretização de cada modelo deve ser analisada individualmente. Portanto, como os MFCCs foram quantizados com $2^{11} = 2048$ componentes, uma quantização bem sucedida do espaço de MFCCs produziria uma discretização ideal em que $H(X) = 11$ bits são usados. Após a quantização, usando o modelo treinado, a entropia estimada usando a sequência S_X foi $\hat{H}(X) = 10{,}85$ bits, mostrando que a discretização dos MFCCs foi bem sucedida e está usando efetivamente o espaço (i.e., as amostras ocupam as diferentes partes do espaço particionado de maneira aproximadamente uniforme). Para essa estimação de entropia, o estimador empírico foi utilizado, pois, como não há a redução do espaço amostral devido ao condicionamento a um outro evento, o tamanho da amostra é grande o suficiente para isso.

O espaço da prosódia foi quantizado com $2^9 = 512$ componentes. Logo, a entropia da variável Y_{test} deve ser de aproximadamente $H(Y_{\text{test}}) = 9$ bits. Desse modo, a discretização obtida com o modelo está satisfatória por possuir a entropia empírica de $\hat{H}(Y_{\text{test}}) = 8{,}98$ bits. A combinação dessa informação com a da quantização mostra que os dois tipos de símbolos estão ocupando efetivamente o particionamento do espaço criado pelos modelos. No entanto, o

compartilhamento do espaço entre esses símbolos ainda precisa ser investigado.

Se o compartilhamento de informação do espaço entre os símbolos não existir, a entropia conjunta $H(X, Y_{\text{test}})$ será a soma $H(X) + H(Y_{\text{test}})$ [1]. A consequência desse resultado seria que, apesar dos dois símbolos serem produzidos simultaneamente, eles estariam condicionados a fenômenos independentes entre si. Se isso acontecesse, a independência entre as duas características para todos os oradores já estaria sendo estabelecida nessa base e o procedimento do teste de hipótese nula seria trivial. No entanto, a entropia conjunta estimada foi $\hat{H}(X, Y_{\text{test}}) = 15{,}58$ bits, o que indica que há uma informação mútua $\hat{I}(X, Y_{\text{test}}) = 4{,}25$ bits entre as duas sequências [2]. Em termos relativos, a razão $\hat{I}(X, Y_{\text{test}})/\hat{H}(X, Y_{\text{test}})$ demonstra que aproximadamente 27% de toda informação contida do espaço compartilhado é redundante. Porém, a significância estatística dessa dependência entre as sequências ainda precisa ser aferida com os testes de hipótese nula para cada orador, que foram conduzidos conforme o procedimento detalhado no Capítulo 3.

Os resultados dos testes de hipótese para cada orador, apresentados na Tabela 1, permitiram cumprir o objetivo específico de determinar a relevância estatística da dependência entre as características após quantificá-la via entropia condicional. Cada teste avaliou a relevância estatística da hipótese nula H_0 das sequências S_{X_j} e $S_{Y_{j,\text{test}}}$ referentes ao j-ésimo orador serem independentes com o $\alpha = 0{,}05$. Os valores-p dos testes foram obtidos empiricamente como menores ou iguais a 10^{-4} uma vez que os valores de cardinalidade efetiva com entropia condicional $C_{Y_j|X_j,\text{test}}$ foram inferiores ao valor mínimo de cardinalidade $C_{Y_j|X_j}$ obtido com as sequências com símbolos permutados, o que possibilita a determinação de um limite superior para o verdadeiro valor-p. Além disso, os resultados também demonstraram que a cardinalidade efetiva entre as características variou para cada orador, o que provavelmente significa que o acoplamento de informação prosódica nos MFCCs pode acontecer em diferentes intensidades. Mas, apesar da variação dessas intensidades, alguns elementos podem ser destacados como:

1. a concentração da cardinalidade ocorre ao redor de 20 símbolos para oradores do sexo feminino (com índice menor ou igual a vinte).

2. a concentração da cardinalidade está em média ao redor de 30 símbolos para oradores masculinos.

3. a cardinalidade efetiva média de todos os oradores atinge aproximadamente 27,07 símbolos (valor associado a uma entropia condicional de 4,76 bits, que representa a incerteza residual em Y_{test} após a redução de aproximadamente 4,25 bits proporcionada pela informação mútua $I(X, Y_{\text{test}})$).

---

[1] Isso remete a uma propriedade conhecida da teoria da informação. A de que a entropia conjunta é aditiva para fontes independentes.

[2] A informação mútua de duas variáveis aleatórias, dada por $I(X, Y_{\text{test}}) = H(X) + H(Y_{\text{test}}) - H(X, Y_{\text{test}})$, é a medida da dependência mútua entre as duas variáveis. Mais especificamente, a informação mútua quantifica a informação que uma variável aleatória contém acerca da outra.

Tabela 1 – Cardinalidades $C_{Y_j|X_j,\text{test}}$ obtidas no teste de cada orador, parâmetros da distribuição empírica de $C_{Y_j|X_j}$ (estimados com 10000 permutações) e o resultado do teste com $\alpha = 0,05$.

Oradores	**Cardinalidade**	**Mediana**	**Mínimo**	**Máximo**	**valor-p**	**Resultado**
1	23,80	65,85	59,80	72,30	$\leq 1 \times 10^{-4}$	rejeição
3	20,10	67,25	61,10	72,60	$\leq 1 \times 10^{-4}$	rejeição
9	22,90	59,50	54,00	65,50	$\leq 1 \times 10^{-4}$	rejeição
10	23,80	74,55	67,80	81,80	$\leq 1 \times 10^{-4}$	rejeição
13	22,00	67,35	61,30	73,10	$\leq 1 \times 10^{-4}$	rejeição
14	24,60	57,30	54,30	60,40	$\leq 1 \times 10^{-4}$	rejeição
15	20,10	51,55	48,60	55,00	$\leq 1 \times 10^{-4}$	rejeição
17	22,00	58,50	52,90	64,20	$\leq 1 \times 10^{-4}$	rejeição
18	25,30	69,20	64,30	74,80	$\leq 1 \times 10^{-4}$	rejeição
19	20,70	52,00	47,30	58,20	$\leq 1 \times 10^{-4}$	rejeição
21	28,60	66,70	60,40	73,40	$\leq 1 \times 10^{-4}$	rejeição
22	31,60	67,60	61,40	74,00	$\leq 1 \times 10^{-4}$	rejeição
26	33,00	64,45	58,60	71,40	$\leq 1 \times 10^{-4}$	rejeição
27	33,00	83,90	76,50	90,80	$\leq 1 \times 10^{-4}$	rejeição
29	31,90	71,00	65,20	77,70	$\leq 1 \times 10^{-4}$	rejeição
30	31,10	63,15	56,50	69,70	$\leq 1 \times 10^{-4}$	rejeição
33	31,70	61,20	55,00	67,80	$\leq 1 \times 10^{-4}$	rejeição
34	31,20	72,50	66,20	78,80	$\leq 1 \times 10^{-4}$	rejeição
35	31,30	70,45	63,90	76,60	$\leq 1 \times 10^{-4}$	rejeição
38	32,60	69,60	62,70	77,00	$\leq 1 \times 10^{-4}$	rejeição

A rejeição unânime da hipótese nula por todos os oradores indica que metodologias para estimar a F0 a partir dos MFCCs são promissoras. Afinal, aproximadamente 27% da informação contida nos pares de símbolos é compartilhada, e as cardinalidades efetivas mostram que esse acoplamento ocorre nas falas de todos oradores testados. Contudo, a variação desse acoplamento entre os diferentes indivíduos sugere que algoritmos de regressão ou de redes neurais projetados para essa tarefa devem incorporar mecanismos de adaptação ao orador. Essa adaptação é justificada pelo desconhecimento sobre quais elementos determinam os níveis de acoplamento entre as características, visto que alguns podem não ser compartilhados por todo grupo de oradores. Por exemplo, os valores de cardinalidade serem menores no caso feminino indicam que o acoplamento entre as duas características é maior nesse gênero. Essa diferença pode ser explicada pela F0 ser intrinsecamente mais elevada na voz feminina, o que faz com que os harmônicos captados pelos filtros de banda dos MFCCs fiquem mais espaçados. Consequentemente, as variações de F0 associam-se a padrões de MFCCs com menor ambiguidade.

Esses destaques apontam um caminho promissor para estudos futuros sobre o acoplamento de informação espectral-prosódica nos MFCCs. O estudo sobre esse acoplamento no presente trabalho alcançou seu objetivo ao demonstrar que ele existe nos MFCCs do conjunto de dados utilizado, dado que 100% dos testes rejeitaram a hipótese nula de independência. No entanto, os valores obtidos nos testes dificilmente podem ser extrapolados para cenários gravados sob outras

condições (como com diferentes idiomas, ambientes ruidosos, variações de equipamento ou enunciações emotivas), uma vez que a intensidade do acoplamento e seus fatores subjacentes não foram avaliados. Além disso, o tamanho do conjunto empregado pode ter sido insuficiente para capturar a real estrutura da dependência entre as características. Contudo, tais limitações não invalidam este trabalho como uma prova de conceito robusta sobre a existência dessa dependência e a relevância dela, por meio de testes de hipótese.

Após o sucesso dos testes de hipóteses presentes na tabela 1, um elemento mais prático foi abordado para a conclusão deste trabalho: uma possível aplicação de processamento de fala que evidencia a dependência entre os MFCCs e a prosódia. Nesse sentido, o reconhecimento de orador foi escolhido como aplicação por utilizar informação dos dois tipos de características. Mais especificamente, a reprodução do trabalho realizado em Adami (2007) foi escolhida para avaliar o desempenho dos modelos treinados em distinguir entre os oradores testados. Assim, investigou-se se a redundância informacional entre as características afeta de alguma forma os sistemas biométricos na prática.

4.2 Experimento de reconhecimento de orador

A implementação dos sistemas biométricos permitiu a concretização do objetivo específico de avaliar o desempenho das características em um experimento de reconhecimento de orador. Nesses sistemas, as quantizações dos MFCCs e dos vetores prosódicos, semelhantes às realizadas na seção anterior, foram usadas para reconhecer oradores, utilizando o logaritmo da razão de verossimilhança como métrica. As pontuações obtidas nesse processo foram fundidas posteriormente por meio de uma rede neural para avaliar a complementaridade da informação de identidade nas características.

O logaritmo da razão de verossimilhança é a métrica utilizada pelo modelo de fundo universal (UBM) [3] para identificar se uma fala pertence a um determinado orador (REYNOLDS; QUATIERI; DUNN, 2000). A fala representada pelo conjunto de vetores de características $\mathbf{V}$ é avaliada como pertencente ao orador j através do logaritmo da razão de verossimilhança $\Lambda(\mathbf{V})$ (também referido como "LLR") que é descrito pela equação

$$\Lambda(\mathbf{V}) = \log_{10}(p_j(\mathbf{V})) - \log_{10}(p_{ubm}(\mathbf{V})), \tag{4.1}$$

em que $p_j(\mathbf{V})$ é a verossimilhança de $\mathbf{V}$ ao modelo do j-ésimo orador e $p_{ubm}(\mathbf{V})$ é a verossimilhança no modelo UBM. Quando $\Lambda(\mathbf{V}) > \Lambda_{limiar}$ a fala é reconhecida como pertencente ao j-ésimo orador.

[3] Modelo baseado em GMM, no contexto de reconhecimento de orador, que atua na razão de verossimilhança como um referencial neutro e independente dos oradores cadastrados. É implementado com a intenção de captar um universo de manifestações acústicas por meio de um grande volume de dados e muitas componentes Gaussianas (REYNOLDS; QUATIERI; DUNN, 2000).

A verossimilhança $p_j(\mathbf{V})$ é obtida através de uma adaptação de uma cópia do UBM (cujo treinamento foi feito com as amostras do conjunto X_{treino}) ao conjunto de vetores do j-ésimo orador $\mathbf{V}_j = \{v_j^{(i)}\}_{i=1}^{N_j}$, em que $v_j^{(i)}$ é o vetor de característica associado ao i-ésimo quadro de áudio do j-ésimo orador. A primeira etapa dessa adaptação é semelhante ao algoritmo EM (BEIGI, 2011). Especificamente, o conjunto $\mathbf{V}_j$ é utilizado para obter dois valores estatísticos relacionados a componente c do modelo pelas equações

$$i_c = \sum_{i=1}^{N_j} \gamma_c(\mathbf{v}_j^{(i)}) \tag{4.2}$$

e

$$E_c(\mathbf{V}_j) = \frac{1}{i_c} \sum_{i=1}^{N_j} \gamma_c(\mathbf{v}_j^{(i)}) \mathbf{v}_j^{(i)}. \tag{4.3}$$

Os valores estatísticos i_c e $E_c(\mathbf{V}_j)$ são utilizados para calcular os parâmetros adaptados da componente c (BEIGI, 2011). Os parâmetros finais do modelo para o j-ésimo orador são descritos pelas equações

$$\pi_{c,j} = [\alpha_c i_c / N_j + (1 - \alpha_c)\pi_c]\beta, \tag{4.4}$$

$$\boldsymbol{\mu}_{c,j} = \alpha_c E_c(\mathbf{V}_j) + (1 - \alpha_c)\boldsymbol{\mu}_c \tag{4.5}$$

e

$$\boldsymbol{\Sigma}_{c,j} = \alpha_c E_c(\mathbf{V}_j)^2 + (1 - \alpha_c)(\boldsymbol{\Sigma}_c + \boldsymbol{\mu}_c^2) - \boldsymbol{\mu}_{c,j}^2 \ , \tag{4.6}$$

em que $\alpha_c = \frac{i_c}{i_c + 16}$ e β é um fator para garantir que a soma dos pesos de cada componente da mistura seja 1. Esses parâmetros foram calculados para cada orador do conjunto de teste da seção passada.

Diferentemente da seção anterior, os dados de cada orador do conjunto X_{teste} não puderam ser integralmente utilizados para o teste dessa seção, visto que parte do conjunto de vetores deve ser dedicada à adaptação do modelo. Como esse conjunto está associado a 40 falas com duração média inferior a 2 minutos, optou-se por destinar os vetores de $F_{tr,j} = 35$ falas para a adaptação ao j-ésimo orador, reservando os vetores de $F_{te,j} = 5$ falas para os testes. Por conseguinte, o experimento de reconhecimento totalizou 100 testes genuínos (5 falas $\times$ 20 oradores cadastrados) e 1.900 testes de impostores (5 falas $\times$ 19 impostores $\times$ 20 oradores cadastrados).

Os resultados de verossimilhança desses testes devem ser comparados com cautela, visto que os modelos de cada orador podem apresentar variações de viés e escala capazes de comprometer o desempenho global do sistema (REYNOLDS; QUATIERI; DUNN, 2000). Uma solução consolidada para essa limitação é a normalização de LLR, conhecida na literatura como normalização de teste (referida pelo termo inglês "T-norm") (AUCKENTHALER; CAREY; LLOYD-THOMAS, 2000). A pontuação normalizada, $\Lambda^{Tnorm}(\mathbf{V})$, é definida por:

$$\Lambda^{Tnorm}(\mathbf{V}) = \frac{\Lambda(\mathbf{V}) - \mu_{impostor}}{\sigma_{impostor}}, \tag{4.7}$$

em que $\mu_{impostor}$ e $\sigma_{impostor}$ representam, respectivamente, a média e o desvio padrão das pontuações obtidas a partir da avaliação do vetor de teste **V** frente ao conjunto de modelos de impostores. Dessa forma, torna-se possível estabelecer um limiar de decisão único, Λ_{limiar}, sem que a acurácia do sistema seja prejudicada por variações entre modelos.

O procedimento padrão para a aplicação da T-norm consiste no emprego de um conjunto de modelos de oradores impostores obtidos a partir de uma base de dados externa para maximizar a robustez em cenários reais (AUCKENTHALER; CAREY; LLOYD-THOMAS, 2000). Todavia, visto que o escopo desta seção não abrange o desenvolvimento de uma aplicação comercial, adotou-se uma abordagem alternativa: a normalização foi realizada utilizando-se modelos de oradores neutros que não correspondem à identidade alegada nem ao locutor da fala de teste corrente. Embora essa estratégia de exclusão dinâmica possa impor restrições logísticas em sistemas de produção, ela se mostra plenamente suficiente para o escopo desta seção, cujo objetivo principal é analisar estritamente a informação discriminativa contida nos modelos.

No trabalho em Adami (2007) três modelos de UBM foram utilizados, em que dois são equivalentes aos modelos treinados na seção anterior. Modelos com a mesma quantidade de componentes Gaussianas que foram treinados com vetores prosódicos ou de MFCCs obtidos de oradores excluídos da etapa de teste. Além desses modelos, um terceiro modelo foi utilizado, um modelo não-paramétrico treinado com símbolos de uma codificação específica capaz de capturar informação da dinâmica prosódica. Assim, o desempenho desse terceiro modelo não está diretamente relacionado com o escopo geral deste trabalho.

O desempenho do modelo com símbolos da dinâmica prosódica é obtido por meio da avaliação da pontuação final do modelo $\Lambda_{j,\text{símbolos}}$ por seu limiar específico $\Lambda_{\text{limiar}}^{\text{símbolos}}$. O mesmo procedimento é realizado com os outros dois modelos: a pontuação com vetores prosódicos $\Lambda_{j,\text{prosódia}}$ tem o limiar $\Lambda_{\text{limiar}}^{\text{prosódia}}$ e a pontuação com vetores de MFCCs $\Lambda_{j,\text{MFCCs}}$ possui o limiar $\Lambda_{\text{limiar}}^{\text{MFCCs}}$. Estes três limiares são estabelecidos pelo ponto de EER (do inglês, “Equal Error Rate”). Este é um ponto de compromisso entre a taxa de falsa aceitação (‘False Acceptance Rate’ - FAR) e a taxa de falsa rejeição (‘False Rejection Rate’ - FRR) (ADAMI, 2007). Ele é determinado pelo ponto mínimo no gráfico de compromisso de erro de detecção (‘Detection Error Tradeoff’ - DET). Nesse gráfico, os sistemas com ótimos desempenhos possuem curvas próximas à origem (MARTIN et al., 1997).

A análise da Figura 4 revela que o sistema baseado em vetores de MFCC (linha verde) apresenta o melhor desempenho global absoluto, atingindo um EER de 4,00%, o que demonstra uma alta capacidade de discriminação e robustez. Em segundo lugar, o sistema de símbolos prosódicos (linha vermelha) exibe um desempenho intermediário muito sólido, registrando um EER de 7,00%. Por fim, o sistema de vetores prosódicos (linha magenta) apresenta o desempenho mais modesto entre os três avaliados, com um EER de 25,00%. No entanto, vale destacar que este último sistema operou substancialmente acima do patamar de um classificador puramente aleatório (cujo EER teórico seria de 50%), evidenciando que todos os métodos implementados

extraíram com sucesso padrões biométricos da fala.

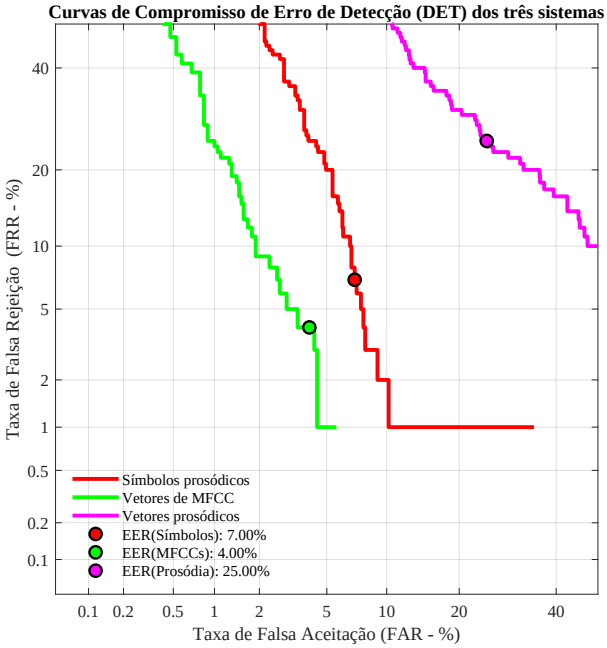


Figura 4 – Curvas de DET para os três sistemas implementados com destaque aos pontos de EER deles.

Os resultados da Figura 4 confirmam o sucesso na reprodução dos sistemas de Adami (2007), mesmo sob condições de restrição de dados, uma vez que a hierarquia de eficácia das características foi fielmente mantida. O EER de 25% observado para os vetores prosódicos — em comparação aos 15,2% (no conjunto NIST-SRE 2001) e 17,1% (no conjunto NIST-SRE 2003) do trabalho original — reflete uma degradação esperada devido ao volume limitado de áudios para a adaptação dos oradores na base utilizada. Contudo, esse mesmo fator justifica o recuo no desempenho dos MFCCs para 4%, ficando pior que os 0,7% (NIST-SRE 2001) e 2,5% (NIST-SRE 2003) da referência. Por outro lado, a limitação de variabilidade fonética do conjunto de dados propiciou uma melhora nos símbolos prosódicos, cujo EER de 7% superou os 11,4% (NIST-SRE 2001) e 14,2% (NIST-SRE 2003). Portanto, apesar das diferenças de desempenho, os três sistemas implementados estão operando conforme esperado.

A etapa seguinte na reprodução do trabalho de Adami (2007) consistiu na realização da fusão dos sistemas. A principal justificativa para essa abordagem reside na complementaridade de informações entre as características avaliadas, que seria esperada caso elas fossem independentes. Na literatura, destaca-se que diferentes elementos da fala, quando combinados, criam um perfil biométrico mais robusto: enquanto os MFCCs capturam a anatomia do trato vocal, os extratores prosódicos modelam o estilo comportamental do orador. A fusão permitiria que essas dimensões distintas fossem analisadas simultaneamente. Caso operassem de maneira verdadeiramente complementar, a combinação dessas fontes preencheria lacunas na descrição individual de cada característica. Esta combinação é feita no trabalho original de Adami (2007) com a fusão por

níveis de verossimilhança, uma técnica consolidada que preserva a interpretação estatística de cada modelo e realiza um mapeamento direto dela para um espaço de decisão comum (CAMPBELL; REYNOLDS; DUNN, 2003).

A fusão por verossimilhança é baseada em uma rede neural densa linear com "one-hot encoding" e ponderação de classe para lidar com o desbalanceamento dos dados (CAMPBELL; REYNOLDS; DUNN, 2003). A rede neural densa linear é um modelo que realiza o mapeamento de um vetor de entrada $\mathbf{v_1} \in \mathbb{R}^{d_e}$ para um vetor de saída $\mathbf{v_2} \in \mathbb{R}^{d_s}$ por meio de uma transformação afim dada por $\mathbf{v_2} = \mathbf{W}\mathbf{v_1} + \mathbf{b}$ (MURPHY, 2022). Nessa formulação, a variável d_e que denota a dimensão do espaço de entrada da rede foi determinada pela quantidade de diferentes modelos sendo fundidos. Ao todo, quatro redes de fusão foram treinadas, uma para cada par de diferentes modelos sendo combinados com $d_e = 2$ e uma combinando os três modelos com $d_e = 3$. Além disso, a dimensão do espaço de saída dessas redes foi $d_s = 2$ por causa das duas classes do "one-hot encoding", uma para indicar identificação e outra para negá-la. Desse modo, cada rede para fusão por verossimilhança foi representada pelo conjunto de pesos e vieses $\mathbf{W} = \{\{\alpha_0^b, \alpha_w^b\}_{w=1}^{d_e}\}_{b=1}^{d_s}$, em que α_0^b é um viés para o valor da saída b e α_w^b é a ponderação da entrada w para a saída b. Com isso, essas redes conseguem ponderar os valores de LLR dos modelos fundidos para fornecer uma decisão.

O treinamento da rede é feito utilizando um subconjunto de LLR associados às 35 falas que foram utilizadas para adaptar os modelos a cada j-ésimo orador. Esse subconjunto é formado dos LLR das 32 falas que não foram selecionadas para o subconjunto de validação. A escolha de 3 falas aleatórias para a validação, aproximadamente 10% das falas, mostrou-se importante porque os LLRs correspondentes permitiram avaliar o treinamento da rede com dados externos ao treinamento. Cada fala é associada aos LLRs fornecidos pelos 20 modelos cadastrados, sendo considerada legítima em um deles e impostora nos outros 19. Com isso, no treinamento, o número total de testes legítimos é $T_1 = 640$ e o número de testes de impostores é $T_0 = 12160$. No entanto, antes de alimentar esses $T = T_0 + T_1 = 12800$ valores na entrada da rede é conveniente normalizá-los. Esta prática é chamada de padronização ("standardization") no treinamento de redes neurais (MURPHY, 2022). Ela é feita através dos valores de $\mu_{\text{treinamento}}$ e $\Sigma_{\text{treinamento}}$ que são extraídos dos dados de treinamento. Esses valores de média e desvio padrão devem ser salvos junto com o modelo treinado para normalizar os LLRs na etapa de validação e de testes.

Outro procedimento padrão da literatura é o uso da função de "Softmax" em redes com "one-hot encoding" na saída (MURPHY, 2022). O rótulo da saída b dessa rede é f_b que é aproximado pelo valor $\hat{f}_b$ na saída da rede pela equação

$$\hat{f}_b = \frac{\exp z_b}{\sum_{b=1}^{d_s} \exp z_b}, \tag{4.8}$$

em que z_b é a combinação linear dos pesos da rede para a saída b, e d_s é o número total de

classes do modelo. Esta combinação linear é descrita pela equação

$$z_b = \alpha_0^b + \sum_{w=1}^{d_e} \alpha_w^b \Lambda(\mathbf{V}_w), \tag{4.9}$$

em que $\mathbf{V}_w$ é o tipo de característica sendo avaliada na entrada w da rede. Esse procedimento faz com que a saída seja aproximada por uma função de densidade probabilística. Isso permite que uma função de divergência como a entropia cruzada seja utilizada para definir o erro entre o rótulo na saída e a estimação na saída da rede.

A função de erro J é a entropia cruzada entre a saída f_b e a estimação $\hat{f}_b$ que é descrita pela equação

$$J = -\sum_{b=1}^{d_s} f_b \log \hat{f}_b. \tag{4.10}$$

O uso do gradiente negativo de J é o procedimento padrão para se treinar uma rede neural para classificação multiclasse (BISHOP; NASRABADI, 2006). Quando a função de erro é a entropia cruzada, ele pode ser facilmente estimado para cada peso ajustado através do gradiente em relação a z_b que é descrito pela equação

$$\frac{\partial J}{\partial z_b} = \hat{f}_b - f_b. \tag{4.11}$$

A regra da cadeia estabelece que $\frac{\partial J}{\partial \alpha} = \frac{\partial J}{\partial z_b} \cdot \frac{\partial z_b}{\partial \alpha}$, assim α_0^b é ajustado pela equação

$$\alpha_0^b = \alpha_0^b - \epsilon_0 \cdot \epsilon_b (\hat{f}_b - f_b) \cdot 1 \tag{4.12}$$

e α_w^b é ajustado com a equação

$$\alpha_w^b = \alpha_w^b - \epsilon \cdot \epsilon_b (\hat{f}_b - f_b) \cdot \Lambda(\mathbf{V}_w), \tag{4.13}$$

em que ϵ é um valor pequeno como 0,001 para ponderar o ajuste a cada iteração e ϵ_b é o peso de classe associado ao rótulo pertencente à saída b (MURPHY, 2022). Esse último peso é uma solução simples da literatura para compensar quando classes estão desbalanceadas, ele é descrito pela equação

$$\epsilon_b = \frac{T}{d_s \cdot T_b}. \tag{4.14}$$

A classificação na saída de cada sistema de fusão foi definida por limiares baseados no ponto de Equal Error Rate (EER). O desempenho das fusões para esse ponto de operação está representado na Figura 5. A combinação entre símbolos e vetores prosódicos apresentou o pior resultado, com taxa de 7,00%, o que equivale ao EER do sistema de símbolos isolado. A fusão entre vetores prosódicos e MFCCs atingiu 3,47%, representando uma melhora de 13% em relação ao uso exclusivo de MFCCs. Os outros dois sistemas obtiveram desempenhos muito próximos: a fusão entre símbolos prosódicos e MFCCs resultou em 2,16%, enquanto a combinação tripla (MFCCs, símbolos e vetores prosódicos) obteve 2,21% — uma diferença

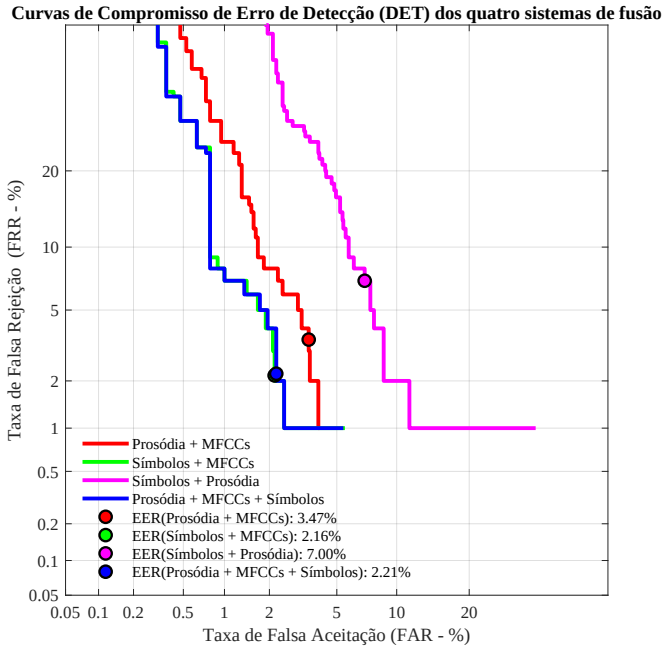


Figura 5 – Curvas de DET para os quatro sistemas de fusão implementados com destaque aos pontos de EER deles.

relativa de aproximadamente 2%. Em síntese, os sistemas de fusão de MFCCs com símbolos prosódicos e vetores prosódicos reduziu o erro em 46% e 13%, respectivamente, ao passo que as outras duas fusões com vetores prosódicos não geraram impacto significativo.

Os resultados na Figura 5 indicam que o sistema de fusão implementado em Adami (2007) foi reproduzido com sucesso, a despeito das discrepâncias decorrentes da escassez de dados já discutidas. A fusão dos sistemas baseados em MFCCs com os de vetores ou símbolos prosódicos produziu taxas de 3,47% e 2,16%, respectivamente, divergindo dos valores de 0,5% (no conjunto NIST-SRE 2001) e 2,3% (no conjunto NIST-SRE 2003) observados na referência. Contudo, a melhora de desempenho de 13% obtida com os vetores prosódicos está na faixa entre 8% e 28,57% de melhora relatada em Adami (2007). Por outro lado, a melhora de 46% na fusão com símbolos prosódicos provavelmente se deve à limitação de variabilidade fonética do conjunto de dados, a qual alavancou o desempenho do sistema com símbolos prosódicos. O desempenho deste último equivale ao resultado de sua fusão com o dos vetores prosódicos (ambos em 7,00%), enquanto na referência houve melhora, 7,2% (no conjunto NIST-SRE 2001) e 10,2% (no conjunto NIST-SRE 2003). Este cenário sugere que a degradação gerada pela falta de dados foi mais acentuada no sistema com vetores prosódicos. Tal fenômeno também justifica o leve declínio na fusão tripla, que registrou com 2,21%; isto é, a inserção do sistema com vetores prosódicos atuou como um ruído marginal, diferentemente do trabalho original, no qual a fusão dos três sistemas apresentou uma leve melhora (0,4% no conjunto NIST-SRE 2001 e 2,2% conjunto NIST-SRE 2003). Em suma, o sistema de fusão foi implementado de maneira bem-sucedida, embora persistam distorções causadas pelas condições de treinamento. O Quadro

2 sintetiza os resultados de todos os sistemas e fusões.

		Conjunto de dados		
		NIST-SRE 2003	NIST-SRE 2001	Ynogutti
Sistemas	Vetores de prosódia	17,1%	15,2%	25,00%
	Vetores de MFCCs	2,5%	0,7%	4,00%
	Símbolos prosódicos	14,2%	11,4%	7,00%
	Prosódia + MFCCs	2,3%	0,5%	3,47%
	Prosódia + Símbolos	10,2%	7,2%	7,00%
	Símbolos + MFCCs	2,3%	0,5%	2,16%
	Prosódia + Símbolos + MFCCs	2,2%	0,4%	2,21%

Quadro 2 – Desempenhos dos sistemas de identificação com diferentes características ou das fusões deles, nos três conjuntos de dados avaliados.

Por fim, o objetivo específico de analisar o impacto da dependência entre as características no reconhecimento de orador foi explorado mediante a fusão dos valores LLR com redes neurais. Essa fusão elucidou a interdependência entre os MFCCs e os vetores prosódicos ao propiciar um ganho marginal de 13,25% na eficácia global. Este incremento reduzido sinaliza uma expressiva sobreposição de informação biométrica entre ambas as frentes, reforçando os achados da seção anterior sobre a presença de informações prosódicas nos MFCCs. Portanto, este acréscimo marginal de desempenho deve ser interpretado mais como reflexo de redundância do que como evidência de complementaridade entre as duas características, contrapondo as premissas de Adami (2007).

5 Conclusão

Neste trabalho, questionou-se a premissa dominante na literatura de que os MFCCs não possuem informação prosódica. Essa premissa assume que, por estarem associados à filtragem do trato vocal, tais coeficientes seriam independentes da prosódia, a qual, por sua vez, é vinculada à fonte glótica (pregas vocais). No entanto, estudos de reconhecimento de orador apontam que os MFCCs carregam informação prosódica de identidade. Diante disso, esse trabalho propôs um procedimento de teste estatístico para avaliar se há dependência entre os MFCCs e um conjunto de características prosódicas (combinação de energia e F0), além de implementar um experimento biométrico para mensurar o impacto prático dessa relação.

O primeiro passo consistiu em modelar o espaço de cada conjunto de características em símbolos por meio de uma quantização vetorial baseada em GMMs. Para tanto, os dados de metade dos oradores foram reservados exclusivamente para o treinamento dos modelos. Os GMMs convergiram em aproximadamente cinco iterações e não apresentaram discrepâncias significativas em relação aos oradores de teste. Contudo, a quantização dos MFCCs com 2048 centróides resultou em uma SNR_Q média baixa (2,6 dB), apesar de essa quantidade de centróides ser considerada satisfatória na literatura de reconhecimento de orador. Apesar disso, as quantizações foram validadas, uma vez que a informação contida nos símbolos de teste se aproximou da informação disponível em cada modelo, apresentando um compartilhamento mútuo de informação de aproximadamente 27%.

Após a confirmação desse compartilhamento, os testes de hipótese foram implementados para cada orador. Neles, os símbolos prosódicos foram permutados para obter uma distribuição empírica da cardinalidade efetiva condicional entre 10 mil permutações e os símbolos de MFCCs. Essas cardinalidades efetivas foram calculadas a partir da entropia condicional entre os pares de sequências, expressando a incerteza sobre os símbolos prosódicos ao se conhecerem os de MFCCs. Se os valores de entropia condicional associados aos pares de símbolo naturais apresentassem uma significância estatística alta, a hipótese nula de independência não poderia ser rejeitada. No entanto, todos os testes rejeitaram a hipótese nula com um $\alpha \leq 1/10000$. Além disso, a cardinalidade efetiva média foi de 27,07 símbolos, com uma média de $\approx$20 símbolos nos testes femininos e $\approx$30 símbolos nos testes masculinos.

Após os testes, os modelos treinados foram reaproveitados em um experimento biométrico com os oradores testados. O desempenho da identificação com MFCCs obteve uma taxa de erro de 4,00% enquanto o uso do conjunto prosódico obteve 25,00%. Esses sistemas foram combinados por meio de uma fusão por verossimilhança com uma rede neural. O desempenho da combinação foi apenas 3,47%, o que sinaliza que a fusão de ambos resultou em um ganho marginal de desempenho (uma melhoria relativa de 13,25% sobre o erro anterior). Esse pequeno

ganho deve ser interpretado como um indicativo de redundância da informação prosódica de identidade presente nos MFCCs, reforçando a dependência estatística encontrada previamente nos testes.

As evidências obtidas neste trabalho devem ser interpretadas sob a ótica de suas delimitações metodológicas como uma prova de conceito. Diante disso, apontam-se as seguintes diretrizes para trabalhos futuros: (i) investigar o impacto da variação do número de componentes Gaussianas nos modelos de quantização; (ii) incluir cenários ruidosos devido a existência do efeito Lombard, que sabidamente influencia a dependência; e (iii) buscar elucidar as causas fisiológicas ou linguísticas subjacentes a esse fenômeno. Por conseguinte, embora a extrapolação dos achados atuais exija prudência devido às condições controladas do conjunto testado, este trabalho consolida uma metodologia sólida para a desmistificação da independência entre os MFCCs e a prosódia.

Referências

ABDUL, Z. K.; AL-TALABANI, A. K. Mel frequency cepstral coefficient and its applications: A review. *Ieee Access*, IEEE, v. 10, p. 122136–122158, 2022.

ADAMI, A. G. Modeling prosodic differences for speaker recognition. *Speech Communication*, v. 49, n. 4, p. 277–291, abr. 2007. ISSN 01676393. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0167639307000271>.

ADAMI, A. G. et al. Modeling prosodic dynamics for speaker recognition. In: IEEE. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* [S.l.], 2003. v. 4, p. IV–788.

AHSAN, M. M.; LUNA, S. A.; SIDDIQUE, Z. Machine-learning-based disease diagnosis: A comprehensive review. In: MDPI. *Healthcare*. [S.l.], 2022. v. 10, n. 3, p. 541.

ALAMRI, F. S. et al. Investigating the impact of combined spectral and prosodic features on speech emotion recognition. *IEEE Access*, IEEE, v. 14, p. 5719–5732, 2026.

ALCAIM, A.; SOLEWICZ, J. A.; MORAES, J. A. de. Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no rio de janeiro. *Journal of Communication and Information Systems*, v. 7, n. 1, 1992.

ALLEN, G. D. Acoustic level and vocal effort as cues for the loudness of speech. *The Journal of the Acoustical Society of America*, AIP Publishing, v. 49, n. 6B, p. 1831–1831, 1971.

ANANTHAKRISHNAN, S.; NARAYANAN, S. S. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE transactions on audio, speech, and language processing*, IEEE, v. 16, n. 1, p. 216–228, 2008.

ASHBY, M.; MAIDMENT, J. *Introducing Phonetic Science*. [S.l.]: Cambridge University Press, 2005. (Cambridge Introductions to Language and Linguistics).

ASHOUR, G.; GATH, I. Characterization of speech during imitation. In: *EUROSPEECH*. [S.l.: s.n.], 1999. p. 1187–1190.

ATAL, B. S. Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, v. 52, n. 6B, p. 1687–1697, 1972. Publisher: Acoustical Society of America.

ATAL, B. S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *the Journal of the Acoustical Society of America*, Acoustical Society of America, v. 55, n. 6, p. 1304–1312, 1974.

AUCKENTHALER, R.; CAREY, M.; LLOYD-THOMAS, H. Score normalization for text-independent speaker verification systems. *Digital signal processing*, Elsevier, v. 10, n. 1-3, p. 42–54, 2000.

BEIGI, H. *Fundamentals of Speaker Recognition*. [S.l.]: Springer Science & Business Media, 2011.

BIAU, D. J.; JOLLES, B. M.; PORCHER, R. P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics and Related Research®*, LWW, v. 468, n. 3, p. 885–892, 2010.

BISHOP, C. M.; NASRABADI, N. M. *Pattern recognition and machine learning*. [S.l.]: Springer, 2006. v. 4.

BROOKES, M. *VOICEBOX: Speech Processing Toolbox for MATLAB*. 1997. Acessado em: 26 jul. 2026. Disponível em: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>.

BRUMM, H.; ZOLLINGER, S. A. The evolution of the lombard effect: 100 years of psychoacoustic research. *Behaviour*, Brill, v. 148, n. 11-13, p. 1173–1198, 2011.

CAMPBELL, J. P.; REYNOLDS, D. A.; DUNN, R. B. Fusing high-and low-level features for speaker recognition. In: *INTERSPEECH*. [S.l.: s.n.], 2003. p. 2665–2668.

CAMPBELL, W. et al. Phonetic speaker recognition with support vector machines. *Advances in neural information processing systems*, v. 16, 2003.

CHAO, A.; SHEN, T.-J. Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environmental and ecological statistics*, Springer, v. 10, n. 4, p. 429–443, 2003.

CHOWDHURY, S. A.; DURRANI, N.; ALI, A. What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis. *Computer Speech & Language*, Elsevier, v. 83, p. 101539, 2024.

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, IEEE, v. 28, n. 4, p. 357–366, 1980.

DEHAK, N.; DUMOUCHEL, P.; KENNY, P. Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech and Language Processing*, v. 15, n. 7, p. 2095–2103, set. 2007. ISSN 1558-7916. Disponível em: <http://ieeexplore.ieee.org/document/4291597/>.

DEHAK, N. et al. Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 19, n. 4, p. 788–798, maio 2011. ISSN 1558-7916, 1558-7924. Disponível em: <http://ieeexplore.ieee.org/document/5545402/>.

DENG, L.; O'SHAUGHNESSY, D. *Speech processing: a dynamic and optimization-oriented approach*. [S.l.]: CRC Press, 2003.

FANT, G. *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations*. The Hague: Mouton, 1960. v. 2. (Description and Analysis of Contemporary Standard Russian, v. 2).

FLANAGAN, J. L. *Speech Analysis Synthesis and Perception*. 2nd. ed. [S.l.]: Springer Berlin, Heidelberg, 1972.

GARAI, S.; SAMUI, S. Advances in small-footprint keyword spotting: A comprehensive review of efficient models and algorithms. *arXiv preprint arXiv:2506.11169*, 2025.

GLASMACHERS, T. Limits of end-to-end learning. In: PMLR. *Asian conference on machine learning*. [S.l.], 2017. p. 17–32.

GOOD, I. J. The population frequencies of species and the estimation of population parameters. *Biometrika*, Oxford University Press, v. 40, n. 3-4, p. 237–264, 1953.

HASIJA, T. et al. Prosodic feature-based discriminatively trained low resource speech recognition system. *Sustainability*, MDPI, v. 14, n. 2, p. 614, 2022.

HAUTAMÄKI, R. G.; KINNUNEN, T. Why did the x-vector system miss a target speaker? impact of acoustic mismatch upon target score on voxceleb data. *arXiv preprint arXiv:2008.04578*, 2020.

HORVITZ, D. G.; THOMPSON, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, Taylor & Francis, v. 47, n. 260, p. 663–685, 1952.

HOSSAIN, M. A.; TRAINI, E.; AMENTA, F. Machine learning approaches to early detection of parkinson's disease using speech analysis technique. *Neurology International*, MDPI, v. 18, n. 5, p. 88, 2026.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. ed. [S.l.]: Online manuscript, 2026.

KIM, H.; PARK, J.-S. Automatic language identification using speech rhythm features for multi-lingual speech recognition. *Applied sciences*, MDPI, v. 10, n. 7, p. 2225, 2020.

KITAMURA, T. Acoustic analysis of imitated voice produced by a professional impersonator. In: *INTERSPEECH*. [S.l.: s.n.], 2008. p. 813–816.

KURDI, M. Z. Integrating acoustic, prosodic, and phonological features for automatic alzheimer's detection. *Frontiers in Aging Neuroscience*, Frontiers Media SA, v. 18, p. 1786269, 2026.

KUWABARA, H.; SAGISAK, Y. Acoustic characteristics of speaker individuality: Control and conversion. *Speech communication*, Elsevier, v. 16, n. 2, p. 165–173, 1995.

LANCKER, D. R. V.; CANTER, G. J. Impairment of voice and face recognition in patients with hemispheric damage. *Brain and cognition*, Elsevier, v. 1, n. 2, p. 185–195, 1982.

LEHMANN, E. L.; ROMANO, J. P. *Testing statistical hypotheses*. [S.l.]: Springer, 2005.

LÓPEZ-ESPEJO, I. et al. Deep spoken keyword spotting: An overview. *IEEE Access*, IEEE, v. 10, p. 4169–4199, 2021.

MARTIN, A. F. et al. The det curve in assessment of detection task performance. In: *Eurospeech*. [S.l.: s.n.], 1997. v. 4, p. 1895–1898.

MARY, L.; YEGNANARAYANA, B. Extraction and representation of prosodic features for language and speaker recognition. *Speech communication*, Elsevier, v. 50, n. 10, p. 782–796, 2008.

MEHRISH, A. et al. A review of deep learning techniques for speech processing. *Information Fusion*, Elsevier, v. 99, p. 101869, 2023.

MILLER, G. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, Free Press, 1955.

MILNER, B.; SHAO, X. Clean speech reconstruction from mfcc vectors and fundamental frequency using an integrated front-end. *Speech Communication*, Elsevier, v. 48, n. 6, p. 697–715, 2006.

MILNER, B.; SHAO, X.; DARCH, J. Fundamental frequency and voicing prediction from mfccs for speech reconstruction from unconstrained speech. In: *Proc. Interspeech 2005*. [S.l.: s.n.], 2005. p. 321–324.

MISSAOUI, I.; LACHIRI, Z. Robust speaker recognition using perceptual stationary wavelet coefficients and prosodic feature in noisy conditions. *IEEE Access*, IEEE, 2025.

MOLLAIE, F. et al. Medical data types in machine learning and their challenges: A comprehensive review. *Intelligence-Based Medicine*, Elsevier, p. 100427, 2026.

MONTALVAO, J.; CANUTO, J.; CARVALHO, E. On the minimum probability of classification error through effective cardinality comparison. *Journal of Communication and Information Systems*, v. 31, n. 1, 2016.

MURPHY, K. P. *Probabilistic machine learning: an introduction*. [S.l.]: MIT press, 2022.

NASSIF, A. B. et al. Speech recognition using deep neural networks: A systematic review. *IEEE access*, IEEE, v. 7, p. 19143–19165, 2019.

OPPENHEIM, A. V.; SCHAFER, R. W. *Discrete-Time Signal Processing*. 3rd. ed. Upper Saddle River, NJ: Pearson, 2010. ISBN 978-0-13-198842-2.

PORTEŠ, D.; HORÁK, A. Learning optimal prosody embedding codebook based on f0 and energy. In: *Proc. Interspeech 2025*. [S.l.: s.n.], 2025. p. 4728–4732.

PROAKIS, J. G.; MANOLAKIS, D. G. *Digital signal processing: Pearson new international edition*. [S.l.]: Pearson Higher Ed, 2013.

QUATIERI, T. *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education, 2008. ISBN 9780132441230. Disponível em: <https://books.google.com.br/books?id=ns7vHuHiEoMC>.

RABINER, L. R.; SCHAFER, R. W. *Introduction to digital speech processing*. [S.l.]: Now Publishers Inc, 2007. v. 1.

RABINER, L. R.; SCHAFER, R. W. *Theory and applications of digital speech processing*. 1st ed. ed. Upper Saddle River: Pearson, 2011. OCLC: ocn476834107. ISBN 978-0-13-603428-5.

REUTER, J. Real-time pitch tracking algorithms in c to test model extraction. *Bachelor thesis, Kiel University, Department of Computer Science*, 2019.

REYNOLDS, D. A.; QUATIERI, T. F.; DUNN, R. B. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, v. 10, n. 1-3, p. 19–41, jan. 2000. ISSN 10512004. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1051200499903615>.

SALIM, S.; SHAHNAWAZUDDIN, S.; AHMAD, W. Automatic speaker verification system for dysarthric speakers using prosodic features and out-of-domain data augmentation. *Applied Acoustics*, v. 210, p. 109412, 2023. ISSN 0003-682X.

SHAN, W. et al. A 510-nw wake-up keyword-spotting chip using serial-fft-based mfcc and binarized depthwise separable cnn in 28-nm cmos. *IEEE Journal of Solid-State Circuits*, IEEE, v. 56, n. 1, p. 151–164, 2020.

SHANNON, C. E. A mathematical theory of communication. *The Bell system technical journal*, Nokia Bell Labs, v. 27, n. 3, p. 379–423, 1948.

SHARMA, R. et al. Milestones in speaker recognition: R. sharma et al. *Artificial Intelligence Review*, Springer, v. 57, n. 3, p. 58, 2024.

SHRIBERG, E. et al. Modeling prosodic feature sequences for speaker recognition. *Speech communication*, Elsevier, v. 46, n. 3-4, p. 455–472, 2005.

SNYDER, D. et al. X-vectors: Robust dnn embeddings for speaker recognition. In: IEEE. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. [S.l.], 2018. p. 5329–5333.

SÖNMEZ, M. K. et al. A lognormal tied mixture model of pitch for prosody-based speaker recognition. Citeseer, 1997.

SOONG, F. K.; ROSENBERG, A. E. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, IEEE, v. 36, n. 6, p. 871–879, 1988.

SOONG, F. K. et al. A vector quantization approach to speaker recognition. *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, v. 10, p. 387–390, 1985. Disponível em: <https://api.semanticscholar.org/CorpusID:273404735>.

TALKIN, D.; KLEIJN, W. B. A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, v. 495, p. 518, 1995.

THOMPSON, N. C. et al. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, v. 10, n. 2, 2020.

TITZE, I. R. Acoustic interpretation of resonant voice. *Journal of Voice*, Elsevier, v. 15, n. 4, p. 519–528, 2001.

TITZE, I. R. Nonlinear source–filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America*, AIP Publishing, v. 123, n. 5, p. 2733–2749, 2008.

TOLEDANO, D. T.; FERNÁNDEZ-GALLEGO, M. P.; LOZANO-DIEZ, A. Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on timit. *PloS one*, Public Library of Science San Francisco, CA USA, v. 13, n. 10, p. e0205355, 2018.

TORRE, I. P. la et al. To bee or not to bee: Estimating more than entropy with biased entropy estimators. *arXiv preprint arXiv:2501.11395*, 2025.

TRACEY, B. et al. Towards interpretable speech biomarkers: exploring mfccs. *Scientific Reports*, Nature Publishing Group UK London, v. 13, n. 1, p. 22787, 2023.

VERMA, V. et al. A novel hybrid model integrating mfcc and acoustic parameters for voice disorder detection. *Scientific Reports*, Nature Publishing Group UK London, v. 13, n. 1, p. 22719, 2023.

WAYLAND, R. *Phonetics: A practical introduction*. [S.l.]: Cambridge University Press, 2018.

YANG, J. A low-power keyword spotting chip with multiplier-free mfcc feature extractor. *IEICE Electronics Express*, The Institute of Electronics, Information and Communication Engineers, v. 23, n. 9, p. 20250008–20250008, 2026.

YNOGUTI, C. A. *Reconhecimento de fala contínua usando modelos ocultos de Markov*. Tese (Doutorado) — [sn], 1999.

ZHENG, N.; LEE, T.; CHING, P.-C. Integration of complementary acoustic features for speaker recognition. *IEEE Signal Processing Letters*, IEEE, v. 14, n. 3, p. 181–184, 2007.