COMPARAÇÃO FORENSE DE LOCUTOR POR MODELOS LINEARES GENERALIZADOS DE VARIABILIDADE ARTICULATÓRIA E VOCAL NA FALA ENCADEADA

Adelino Pinheiro Silva
http://lattes.cnpq.br/8373538496107754 –https://orcid.org/0000-0002-2796-4841
adelino.pinheiro@policiacivil.mg.gov.br
Polícia Civil de Minas Gerais, Belo Horizonte, MG, Brasil

Maria Mendes Cantoni
http://lattes.cnpq.br/7751617715708501 – https://orcid.org/0000-0001-9515-1802
mmcantoni@gmail.com
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil

RESUMO
A variabilidade é parte inerente da fala e se deve a fatores relacionados aos locutores (e.g. sociolinguísticos e pessoais) e linguísticos (e.g. fonético-fonológicos e coarticulatórios). Para uma mesma mensagem e em um mesmo contexto de produção, a variabilidade extrafalante, anatômica ou fisiológica, deve-se à diferença nos tratos vocais e nas rotinas motoras, enquanto a variabilidade intrafalante é biomecânica e se deve a diferenças na execução dos movimentos da fala do indivíduo. Contudo, ainda não é claro qual o papel dos diferentes componentes do trato vocal na classificação de falantes, e apenas alguns estudos usaram fala contínua. Partindo desta questão, o presente estudo tem por objetivo modelar a variabilidade de locutores considerando as estruturas articulatórias e vocais na fala contínua. Projetou-se um procedimento de classificação baseado em um modelo de regressão no qual a variabilidade linguística previsível é removida e os resíduos são usados como entrada para comparação dos locutores. A elaboração do procedimento e os testes subsequentes foram realizados com 18 gravações da base de dados CEFALA-1. Entre os resultados obtidos, destacam-se: (1) A maior parte da variabilidade acústica extrafalante vem de diferenças relacionadas ao sexo do locutor. (2) Tanto as variáveis articulatórias quanto as vocais são relevantes para a classificação dos falantes, sendo que as vocais discretamente superam as articulatórias, quando dispostas em modelos isolados. Entre as limitações do estudo, observamos que o procedimento depende apenas da variabilidade estática abordada, não dinâmica, e não leva em conta a variabilidade consonantal.

Palavras-chave: Comparação forense de locutor; Variabilidade articulatória; Variabilidade vocal; Fala contínua; Modelos Lineares Generalizados.

FORENSIC SPEAKER COMPARISON BY GENERALIZED LINEAR MODELS OF ARTICULATORY AND VOCAL VARIABILITY IN CONNECTED SPEECH

ABSTRACT
Variability is inherent to speech and arises from both speaker-related factors (e.g. sociolinguistic and personal) and linguistic factors (e.g. phonetic-phonological and coarticulatory). For the same message uttered in the same context, between-speaker variability, which can be anatomical or physiological, results from differences in vocal tract structures and motor routines, while within-speaker variability is biomechanical, stemming from variations in an individual's speech execution. Despite this understanding, the specific roles of different vocal tract components in speaker classification remain unclear, and few studies have utilized continuous speech. This study aims to model speaker variability by considering articulatory and vocal structures in continuous speech. We developed a classification procedure based on a regression model that removes part of context variability, using the residuals for speaker comparison. The development of the procedure and subsequent testing were conducted using 18 recordings from the CEFALA-1 database. Key findings include: (1) Most between speaker acoustic variability is attributed to differences related to the speaker's sex, and (2) both articulatory and vocal variables are significant for speaker classification, with vocal variables slightly outperforming articulatory variables in isolated models. Limitations of the study include its focus solely on static variability, excluding dynamic aspects, and the omission of consonant variability.

Keywords: Forensic speaker comparison; Articulatory variability; Vocal variability; Connected speech; Generalized linear models.

DOI: https://doi.org/10.70365/2764-0779.2024.101

Recebido em: 02/09/2024.
Aceito em: 09/10/2024.

1 INTRODUÇÃO

A comunicação oral é uma das bases da interação social e o método primário de transmitir informações. Os humanos desenvolveram a capacidade vocal para codificar e transmitir tanto eventos concretos como conceitos abstratos. A capacidade vocal é altamente plástica, sendo que a mesma pessoa pode voluntariamente alterar características de sua forma habitual de falar (Kreiman *et al*., 2015). Alterações voluntárias ocorrem, por exemplo, na velocidade da fala, na frequência fundamental e no estilo. Por outro lado, a forma de falar pode sofrer alterações de menor controle, como em situações de forte emoção ou doenças do trato vocal (Lavan, 2019).

O sinal acústico produzido pelo trato vocal – i.e., a vocalização – possui uma vasta gama de informações[1] ou fatores (Flanagan, 2008). Dentre essas informações, podem-se citar a mensagem transmitida, a codificação – e.g., idioma e léxico –; a identidade de grupo – e.g., dialeto –; a identidade do falante – e.g., fisiologia, condições metabólicas –; e as condições do falante – e.g., emoção, fadiga, patologia. Em uma situação de interação falada típica, a informação apresenta uma variabilidade limitada pelas condições de contorno da comunicação oral, que são limitações inerentes aos componentes que atuam no processo comunicativo. Dentro destas condições, podem-se citar a dinâmica, a anatomia do trato vocal – como dimensão, velocidade dos movimentos articulatórios – e a eficiência na transmissão da mensagem – o custo comunicativo para alcançar o objetivo de ser compreendido (Furui, 2018).

A variabilidade da informação presente no sinal acústico da fala pode ser estudada por diferentes tipos de recortes. Quando a variabilidade ocorre dentro de um contexto sociocultural, os dialetos podem ser vistos como sistemas linguísticos que refletem as diferentes formas de falar de um grupo ou comunidade. Labov (1972) mostrou empiricamente como fatores sociais – e.g., classe social, etnia e idade – influenciam a variação dialetal de diferentes grupos de indivíduos. As variações dialetais dentro de um grupo de indivíduos são importantes, pois carregam um caráter de consciência linguística e identidade na manutenção e na diferenciação desses dialetos (Chambers, 1995).

O traço sociolinguístico de um dialeto pode se manifestar no plano acústico. Um grupo pode ter como recorrência um conjunto de processos fonológicos ou a escolha de palavras características. Essas recorrências, que são escolhas do falante dentro da identidade de grupo, influenciam diretamente o sinal acústico e, por sua vez, as medidas realizadas na voz.

[1] No presente trabalho, o conceito de informação refere-se à definição presente na teoria matemática da informação proposta originalmente por Shannon (1948).

Diferentes trabalhos indicam que a frequência de uso de elementos dialetais pode contribuir na identificação de autoria e de falantes (Doddington, 2001; Ishihara, 2021).

Além das especificidades determinadas pelos traços sociolinguísticos, existem diferenças no sinal acústico relacionadas à anatomia e à fisiologia do falante (Fant, 1971; Flanagan, 2013). A variabilidade anatômica ocorre principalmente devido às limitações mecânicas do trato vocal (e.g., dimensões, elasticidade muscular), enquanto a variabilidade fisiológica está relacionada à regulação do organismo, como a energia empregada e a viscosidade do ar no interior do trato vocal. Um dos desafios da comparação forense de locutor (CFL) é compreender e utilizar tais variáveis dentro do modelo estatístico, estabelecendo uma estrutura capaz de isolar as variáveis interferentes e aproveitar os fatores discriminativos no modelo estatístico (Drygajlo, 2019).

Mesmo com as limitações decorrentes das condições de contorno, dois tipos de variabilidade do sinal acústico da fala são classicamente modelados, a intra e a extrafalante (Kilbourn-Ceron & Goldrick 2021). A variabilidade intrafalante é aquela presente na fala de um mesmo locutor, devido às diferenças de como os movimentos da fala são articulados por ele. A variação extrafalante é a variabilidade observada entre falantes distintos devido a diferentes tratos vocais e habilidades motoras.

As variabilidades intra e extrafalante podem ser medidas por meio de diferentes características do sinal acústico, por exemplo, características vocais e articulatórias. Uma característica vocal é uma medida acústica diretamente influenciada pelo ajuste das pregas vocais, e.g., frequência fundamental. Uma característica articulatória refere-se a uma medida influenciada pelo movimento do trato oral, e.g., frequência dos formantes. Essa divisão entre medidas vocal e articulatória deriva do Modelo Fonte-Filtro proposto por Fant (1971).

Conhecendo a variabilidade intra e extrafalante, é possível realizar uma inferência entre semelhança e tipicidade que permite reconhecer um indivíduo pela voz (Kreiman & Sidtis 2011). A evolução da computação digital propiciou o desenvolvimento de métodos para verificação de identidades utilizando a voz como traço biométrico. Paralelamente, o novo paradigma para identificação de indivíduos nas ciências forenses estabeleceu protocolos mais científicos para a atribuição de autoria (Saks & Koehler, 2008). Nesse ponto, a autenticação biométrica e a identificação/verificação forense (especificamente a criminal) se diferem em alguns aspectos.

Na área forense, a comparação de locutor ocorre entre duas fontes (amostras). A primeira, o material questionado, é um vestígio de fato típico

penal e tem origem e autoria desconhecidas. A princípio, o material questionado é arrecadado sem o controle dos parâmetros de qualidade. A segunda é o material padrão de um indivíduo conhecido. Esse material é coletado por livre consentimento de um suspeito, que muitas vezes não deseja ser verificado (que pode acarretar em associação com um típico penal) e pode não ser colaborativo em ceder amostras de voz (Maher, 2009; Silva, 2020).

Os primeiros métodos de comparação de locutor remetem ao trabalho de Kersta (1962), que evoluem para elaboração de perfis de fala a partir de análises auditivas instrumentais (Gfrörer, 2003) até o uso de sistemas baseados em redes neurais profundas (Sztahó & Fejes, 2023). O paradigma atual da identificação/verificação nas ciências forenses recomenda a apresentação quantitativa de resultados na forma de razão de verossimilhança (LR – *likelihood ratio*) calculada em uma comparação estatística dentro de um banco de dados representativo e confiabilidade conhecida (Saks & Koehler, 2008; Morrison, 2009). A aplicação do paradigma para a CFL desperta duas necessidades: (1) a composição dos bancos de dados que representem as raridades e as tipicidades encontradas na população; e (2) a modelagem das características presentes no áudio e na imagem que permitam representar um indivíduo e comparar com os outros modelos, utilizando a razão de verossimilhança como métrica de similaridade/divergência (Campbell *et al*., 2009; Kabir *et al*., 2021).

Uma das formas de aplicação do paradigma da CFL é utilizar modelos paramétricos e processamento padronizado para gerar e comparar fatores discriminativos nos modelos dos indivíduos. Para alguns autores (Gonzalez-Rodriguez et a*l*., 2007), a aplicação de métodos baseados no reconhecimento de padrões agrega à CFL a robustez e a confiabilidade presentes no exame de comparação por perfil molecular (DNA - *deoxyribonucleic acid*). Outro desafio da CFL é compreender e colocar tais fatores discriminativos dentro do modelo estatístico, estabelecendo uma estrutura capaz de isolar os fatores interferentes dos fatores discriminativos (Drygajlo, 2009).

Como colocado anteriormente, o sinal acústico é um conjunto de diferentes informações, e.g. mensagem, dialeto, identidade do falante, condições psicológicas e fisiológicas. Devido a esse acúmulo de informações no sinal acústico, ainda não é claro o papel das diferentes características da voz e do trato vocal (Lee, Keating& Kreiman, 2019) na caracterização da identidade de um locutor. A partir dessa lacuna, o presente trabalho tem por objetivo modelar a variabilidade relacionada ao falante, considerando-se os papéis das estruturas articulatórias e vocais na CFL a partir de fala contínua. As principais perguntas abordadas são: (1) Quanto da variação do falante se deve a

diferenças articulatórias e quanto se deve a diferenças de voz? (2) Quais medidas acústicas são mais robustas para classificação de falantes em fala contínua? Com o intuito de responder tais questões, projetou-se um procedimento de regressão baseado em modelos lineares generalizados (GLM - *generalized linear model*), no qual a variabilidade linguística previsível é removida, e os resíduos são usados como entrada para modelagem dos locutores.

A hipótese do presente trabalho é que o GLM permite remover parte da informação do sinal acústico – como o contexto fonológico – e, com isso, acentuar as demais informações, como a identidade do locutor. Os autores optaram por utilizar uma modelagem estatística baseada em variáveis pragmáticas (no espaço mensurável), devido à transparência do modelo. Diferentemente dos modelos baseados em redes neurais artificiais por variáveis latentes, os modelos estatísticos permitem compreender a influência das variáveis (Mcquisten & Peek, 2009; Wüthrich, 2019) tanto articulatórias quanto vocais.

O principal foco do texto será a apresentação do método de modelagem da variabilidade dos falantes utilizando GLM e a discussão dos resultados. O trabalho justifica-se por contribuir com o desenvolvimento tecnológico da CFL e por contribuir para a compreensão das fontes de variabilidade na execução da fala. Isso posto, a próxima seção descreve a base de dados, os métodos de etiquetagem, as principais medidas acústicas e os *softwares* utilizados. A terceira seção descreve os procedimentos de modelagem da variabilidade, a preparação dos dados, o método de treinamento e validação e a comparação dos locutores. A quarta seção discute as limitações e possibilidades dos resultados, enquanto a última seção resume as principais conclusões e propostas de continuidade.

2 MATERIAIS E MÉTODOS

Neste estudo, foi utilizada uma amostra do *corpus* CEFALA1 (Neto, Silva & Yehia, 2019). A amostra foi selecionada com 18 locutores (oito do sexo feminino e dez do sexo masculino) do mesmo dialeto e foi avaliada a porção contendo fala encadeada (duração média de 2 minutos). Os áudios utilizados foram processados com frequência de amostragem de 16 kHz e 16 bits de profundidade.

A unidade amostral do conjunto de áudio são as vogais e os ditongos, que foram segmentados e rotulados manualmente por pesquisadores treinados. A escolha dessas unidades amostrais é devido ao fato de as vogais e ditongos carregarem componentes tanto vocais quanto articulatórios, e o trato vocal

aberto apresentar um mapeamento menos complexo entre articulação e acústica.

A rotulagem das unidades amostrais buscava codificar os fatores fonológicos:

- Tipo de som, oral ou nasal e se é vogal ou ditongação;
- Contexto sonoro anterior e seguinte, indicando início ou fim de palavra ou os fonemas adjacentes;
- Grau de acento, podendo ser tônico, átono pré-tônico ou átono pós-tônico;
- Número de sílabas da palavra; e
- Posição da vogal em relação à tônica.

Dos fatores fonológicos elencados, foi definido um conjunto de cinco variáveis que podem ser referidas ao contexto em que cada vogal é executada: o sexo do falante, a tonicidade, a posição da sílaba na palavra e os sons anterior e posterior a vogal. O Quadro 1 detalha resumidamente as variáveis referentes ao contexto em que foram utilizadas no presente estudo.

Quadro 1 – Resumo das variáveis de contexto utilizadas no presente estudo.

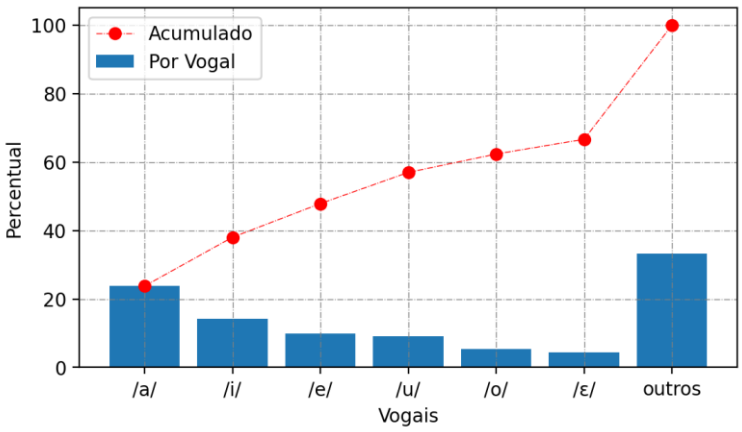
Variável	Descrição
Sexo do falante	Variável categórica que pode assumir os valores feminino ou masculino.
Tonicidade da sílaba	Variável categórica que indica a posição relativa da sílaba em relação à tônica, os valores variam entre tônica, pré-tônica ou pós-tônica.
Posição na palavra	Variável quantitativa discreta que indica em qual sílaba da palavra, a partir do começo, a vogal se encontra. Os valores na amostra variam entre 0 e 6.
Ditongação	Variável dicotômica que indica se a unidade amostral é uma vogal com valor 0 ou ditongo com valor 1.
Fechamento	Variável dicotômica que indica se a sílaba onde se encontra a vogal ou ditongo é fechada por consoante.
Som anterior	Variável categórica que indica o som fonológico que ocorre antes da vogal podendo assumir um valor nulo (início de palavra) ou qualquer vogal (e.g., /a/, /ɛ /, /ĩ/, ...) ou consoante (e.g., /p/, /tʃ/, /n/, ...).
Som posterior	Variável categórica que indica o som fonológico que ocorre depois da vogal podendo assumir um valor nulo (final de palavra) ou qualquer vogal (e.g., /a/, /ɛ /, /ĩ/, ...) ou consoante (e.g., /p/, /tʃ/, /n/, ...).

Fonte: elaborado pelos autores.

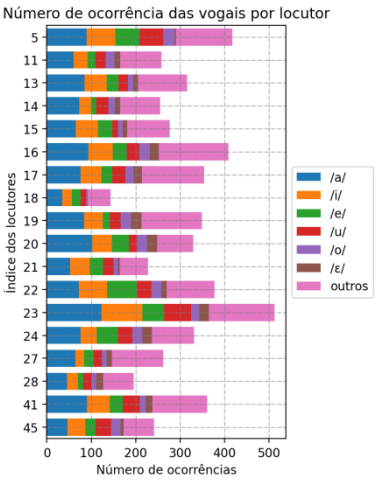
Um total de 5.615 vogais e ditongos foram segmentados – i.e. demarcados nos seus tempos de início e fim no áudio – e rotulados com as variáveis de contexto indicadas no Quadro 1. A amostra apresentou um total de 64 categorias diferentes incluindo as sete vogais orais (/a/, /e/, /ɛ /, /i/, /o/, /ɔ / e /u/), as cinco vogais nasais (/ã/, /ẽ /, /ĩ/, /õ/ e /ũ/) e mais 52 ditongos diversos. A distribuição das diferentes categorias é heterogênea, sendo que 66,7% da amostra apresenta vogais orais mais utilizadas no dialeto de Minas Gerais, e o restante fica distribuído entre as demais variantes. A Figura 1a

apresenta a distribuição das principais ocorrências de vogais na amostra, sendo que as barras verticais indicam as ocorrências percentuais, e a linha vermelha, o percentual acumulado. Na Figura 1b, tem-se a distribuição por locutor. Os gráficos da Figura 1 evidenciam a heterogeneidade presente na amostra tanto na ocorrência das vogais e ditongos quanto no número de amostras por locutor. Os valores absolutos e percentuais por locutor são apresentados na Tabela 2.

Figura 1 - À esquerda, o diagrama de Pereto, no qual as barras verticais indicam o percentual de ocorrência das vogais mais recorrentes no estudo. À direita, a distribuição destas vogais por locutor.



(a) Percentual de ocorrência das unidades amostrais.



(b) Ocorrência das unidades amostrais por locutor.

Fonte: elaborado pelos autores.

Tabela 2 – Distribuição das vogais mais recorrentes utilizadas no estudo indicando entre parênteses o percentual de ocorrência por locutor.

Índice do Locutor	Número de ocorrências das vogais e, entre parênteses, o percentual em relação ao locutor.							
	/a/	/i/	/e/	/u/	/o/	/ɛ /	outras	**Total**
05	90 (21,5)	64 (15,3)	55 (13,2)	53 (12,7)	25 (6,0)	3 (0,7)	128 (30,6)	418 (100,0)
11	60 (23,3)	32 (12,4)	17 (6,6)	24 (9,3)	19 (7,4)	13 (5,0)	93 (36,0)	258 (100,0)
13	85 (26,9)	50 (15,8)	26 (8,2)	22 (7,0)	11 (3,5)	11 (3,5)	111 (35,1)	316 (100,0)
14	73 (28,6)	26 (10,2)	12 (4,7)	28 (11,0)	14 (5,5)	12 (4,7)	90 (35,3)	255 (100,0)
15	65 (23,5)	50 (18,1)	32 (11,6)	13 (4,7)	11 (4,0)	10 (3,6)	96 (34,7)	277 (100,0)
16	94 (23,0)	54 (13,2)	32 (7,8)	28 (6,8)	24 (5,9)	20 (4,9)	157 (38,4)	409 (100,0)
17	76 (21,5)	47 (13,3)	24 (6,8)	30 (8,5)	18 (5,1)	19 (5,4)	140 (39,5)	354 (100,0)
18	35 (24,5)	21 (14,7)	20 (14,0)	13 (9,1)	2 (1,4)	1 (0,7)	51 (35,7)	143 (100,0)
19	84 (24,1)	42 (12,0)	15 (4,3)	25 (7,2)	23 (6,6)	24 (6,9)	136 (39,0)	349 (100,0)
20	102 (31,0)	44 (13,4)	39 (11,9)	18 (5,5)	23 (7,0)	22 (6,7)	81 (24,6)	329 (100,0)
21	52 (22,8)	44 (19,3)	31 (13,6)	23 (10,1)	10 (4,4)	4 (1,8)	64 (28,1)	228 (100,0)
22	72 (19,1)	64 (17,0)	67 (17,8)	32 (8,5)	23 (6,1)	12 (3,2)	107 (28,4)	377 (100,0)
23	123 (24,0)	92 (18,0)	49 (9,6)	61 (11,9)	18 (3,5)	22 (4,3)	147 (28,7)	512 (100,0)
24	76 (23,0)	36 (10,9)	48 (14,5)	32 (9,7)	23 (6,9)	21 (6,3)	95 (28,7)	331 (100,0)

27	64 (24,4)	20 (7,6)	21 (8,0)	19 (7,3)	11 (4,2)	11 (4,2)	116 (44,3)	262 (100,0)
28	46 (23,6)	24 (12,3)	12 (6,2)	18 (9,2)	11 (5,6)	16 (8,2)	68 (34,9)	195 (100,0)
41	91 (25,2)	50 (13,9)	30 (8,3)	38 (10,5)	14 (3,9)	14 (3,9)	124 (34,3)	361 (100,0)
45	47 (19,5)	40 (16,6)	23 (9,5)	34 (14,1)	22 (9,1)	7 (2,9)	68 (28,2)	241 (100,0)
Total	1.335 (23,8)	800 (14,2)	553 (9,8)	511 (9,1)	302 (5,4)	242 (4,3)	1.872 (33,3)	5615 (100,0)

Fonte: elaborado pelos autores.

Para cada uma das unidades amostrais, foram realizadas 26 medidas acústicas classicamente utilizadas em estudos de fala e voz (Garallek, 2022; Kreiman & Sidtis, 2011), sendo elas:

- a duração (tempo em segundos) e a intensidade em decibéis (dB);
- a frequência dos quatro primeiros formantes F1, F2, F3 e F4 e a dispersão dos formantes FD em Hz (Fitch, 1997);
- a frequência fundamental F0 em Hz;
- as amplitudes relativas do primeiro e segundo harmônicos (H1*–H2*) e do segundo e quarto harmônicos (H2*–H4*) em dB; e as inclinações espectrais do quarto harmônico para o harmônico mais próximo de 2 kHz (H4*–H2kHz*) e do harmônico mais próximo de 2 kHz para o harmônico mais próximo de 5 kHz (H2kHz*–H5kHz) em dB/Hz;
- a proeminência do pico cepstral em Hz-1 (CPP) em relação à esperada amplitude derivada por meio de regressão linear (Hillenbrand et al., 1994) e a relação de amplitude entre sub-harmônicos e harmônicos em dB (Shr; Sun, 2002).

Os valores dos harmônicos marcados com "*" foram corrigidos para a influência dos formantes nas amplitudes harmônicas (Hanson & Chuang, 1999; Iseli & Alwan, 2004). As medidas acústicas, com exceção da duração, foram realizadas com a divisão da unidade amostral em quadros de 20 milissegundos e passo de tempo de 5 milissegundos. Das medidas realizadas em cada unidade amostral, foi extraída a média e o coeficiente de variação (CoV – *Coeffient of Variation*) como medida de variabilidade. Não foram avaliadas as variações dinâmicas para a duração (tempo) do som e a intensidade (energia).

As medidas acústicas foram categorizadas como articulatórias quando predominantemente dependentes do movimento do trato vocal; ou como vacais quando predominantemente dependentes da vibração das pregas vocais. Dessa forma, foram classificadas como articulatórias a média e o coeficiente de variação dos quatro primeiros formantes, bem como da dispersão dos formantes. As medidas vocais foram as médias e o coeficiente de F_0, SHR, CPP, H1*-H2*, H2*-H4*, H4*–H2kHz* e H2kHz*–H5kHz. No Quadro 2, é apresentado um resumo das medidas acústicas utilizadas.

A consolidação dos dados foi realizada no formato *tidy data,* de forma que as medidas acústicas, bem como as variáveis de contexto foram dispostas nas colunas, e a unidade amostral (vogal/ditongo) disposta em linhas (Wickham, 2014). Para a análise dos dados, foram utilizados os pacotes estatísticos (*scipy*), de aprendizado de máquina (*scikit-learn*) e de manipulação (pandas) e visualização de dados (*matplotlib*) implementados em *python*.**Quadro 2** – Resumo das medidas acústicas utilizadas no presente estudo indicando a natureza (tipo) da medida, a tendência central, variabilidade e a categoria.

Tipo de medida acústica	Tendência central	Variabilidade	Categoria
1 – Intensidade e duração	Intensidade, Duração	--	Não classificada
2 - Frequência dos formantes	média F_1, média F_2, média F_3, média F_4 e média F_D	Cov F1, Cov F2, Cov F3, Cov F4 e Cov FD	Articulatória
3 - Frequência fundamental	média F_0	Cov F_0	Vocal
4 - Forma espectral da fonte harmônica	média H1*–H2*, média H2*–H4*, média H4*–H2kHz* e média H2kHz*–H5kHz	Cov H1*–H2*, Cov H2*–H4*, Cov H4*–H2kHz* e Cov H2kHz*–H5kHz	Vocal
5 - Ruído espectral/fonte inarmônica	média CPP e média SHR	Cov CPP e Cov SHR	Vocal

Fonte: elaborado pelos autores.

2.1 Análise de Componentes Principais

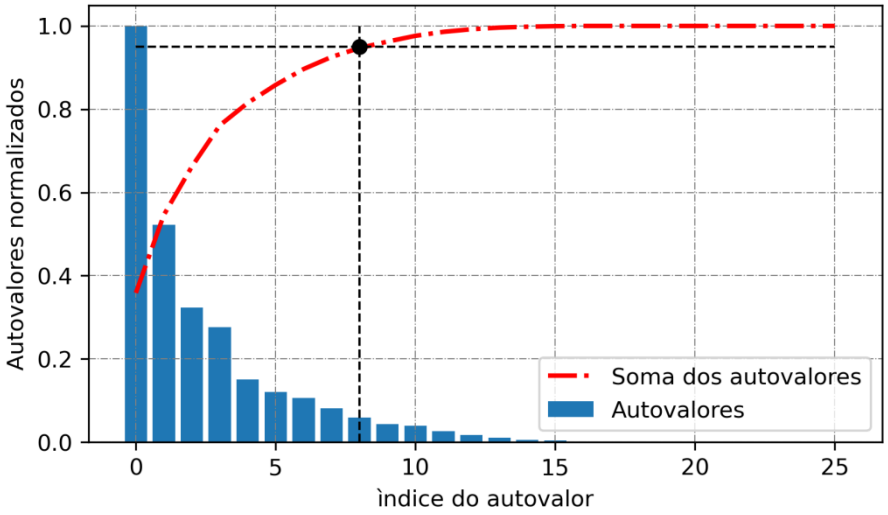
Na análise de dados multidimensionais, é comum que os dados apresentem correlação entre si, mas que mantenham uma relativa independência. No caso de medidas acústicas obtidas do sinal de voz, a literatura relata que diferentes grandezas apresentem uma relação linear na forma de correlação ou não linear na forma de informação mútua (Silva, Vieira & Barbosa, 2019; Lee, Keating & Kreiman, 2019).

Uma técnica de reduzir a correlação entre as variáveis é a análise de componentes principais (PCA - *Principal Component Analysis*). Trata-se de um método não supervisionado de exploração de dados que aplica uma transformação ortogonal linear que projeta o espaço de medições em um espaço de componentes não correlacionadas de acordo com a variabilidade (Duda, Hart & Stork, 2001). A PCA pode ser aplicada na ausência de conhecimento prévio sobre as amostras, agrupando-as com base em similaridades. A partir dos autovalores da matriz de correlação das medidas acústicas, é possível selecionar um número de componentes principais – menor que o original – para construir um novo espaço de análise. Por conseguinte, a análise da variabilidade extralocutor pode ser realizada no espaço das componentes principais, permitindo, inclusive, a realização da comparação entre os locutores tanto no espaço das medidas acústicas quanto no espaço

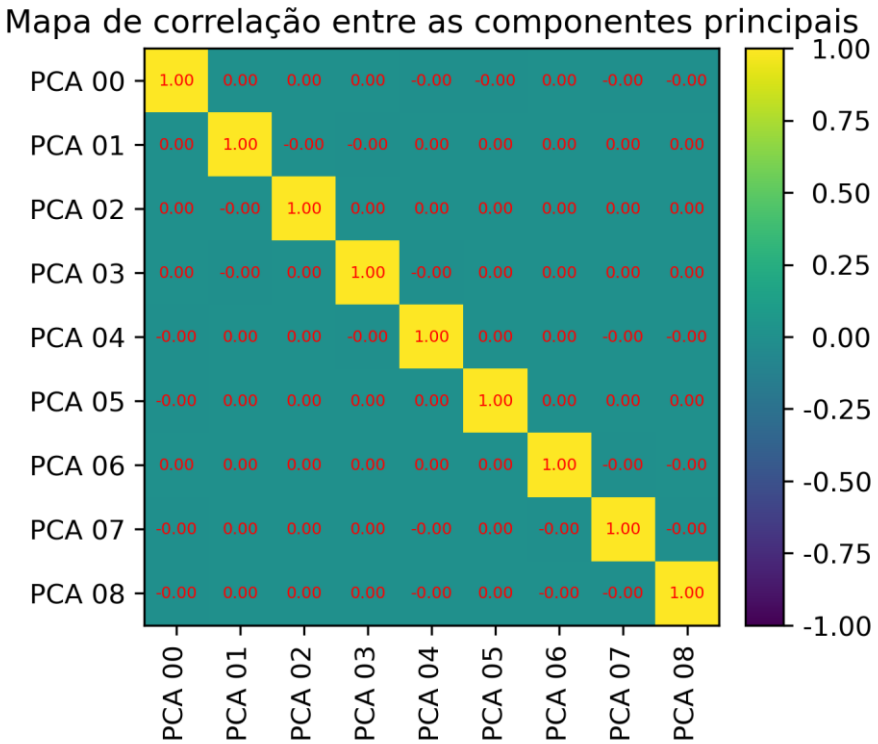
das componentes principais.

Os valores normalizados dos 26 autovalores da matriz de correlação são apresentados na Figura 2a. No gráfico, as barras verticais azuis indicam os valores relativos dos autovalores em relação ao de maior magnitude, enquanto a linha vermelha indica o valor acumulado relativo. O ponto marcado de preto indica índice do autovalor logo acima do valor acumulado de 0,95. Nesse caso, foi realizada a transformação com nove componentes principais. A Figura 2b apresenta o mapa de correlação das nove primeiras componentes principais, indicando uma base ortogonal com correlação cruzada nula.

Figura 2 – À esquerda, o diagrama de Pareto com valores normalizados dos autovalores da matriz de correlação entre as medidas acústicas, indicando que os nove primeiros componentes principais acumulam 95% do valor da soma total. À direita, a matriz de correlação entre as nove primeiras componentes principais.



(a) Autovalores normalizados da matriz de correlação.

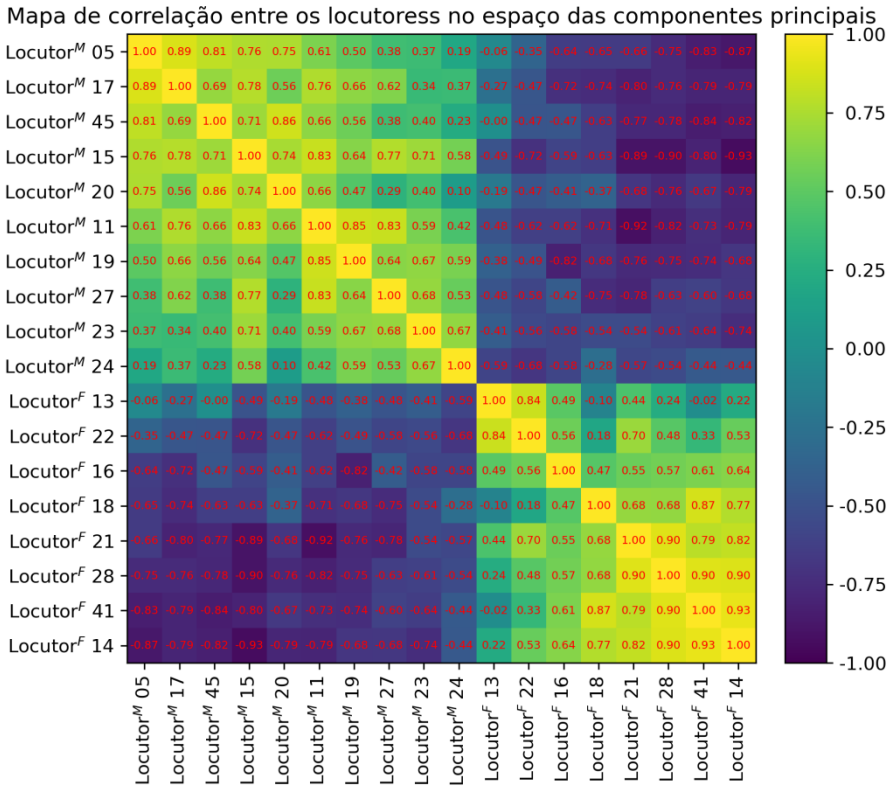


(b) Mapa de correlação entre as nove primeiras componentes principais.

Fonte: elaborado pelos autores.

Após projetar as medidas dos locutores no espaço das componentes principais, notou-se que a principal variabilidade permitia separar os locutores do sexo masculino do sexo feminino. No mapa de correlação entre os locutores da Figura 3, é possível notar os dois grupos. Na etiqueta de marcação, as letras "F" e "M" sobrescritas em cada locutor indicam se os locutores são, respectivamente, do sexo feminino ou masculino.

Figura 3 – Matriz de correlação entre os locutores no espaço das nove primeiras componentes principais, indicando a separação dos locutores "feminino" (porção inferior direita) e "masculino" (porção superior esquerda).



Fonte: elaborado pelos autores.

3 METODOLOGIA DE MODELAGEM E RESULTADOS

3.1 Descrição Procedimental

A aplicação do GLM visa fazer a previsão dos valores das medidas acústicas apresentadas no Quadro 1 de acordo com fatores relativos à variabilidade relativa ao contexto em que um ditongo ou uma vogal é executada.

O GLM considera que as medidas acústicas obtidas Y_n associadas a um locutor n, obtidas na fala encadeada, são oriundas de dois fatores principais. O primeiro é a variabilidade anatômica e fisiológica do falante, e a segunda a variabilidade devido ao contexto acústico.

Dessa forma, o valor Y_n pode ser modelado com influência da variabilidade própria do locutor X_L e uma variabilidade influenciada pelo contexto X_C como

$$Y_n \approx X_{C|n} + X_{L|n} + \varepsilon_n \qquad \text{(Equação 1)}$$

Onde $X_{L|n}$ e $X_{C|n}$ representam, respectivamente, as variabilidades de

locutor e de contexto associadas ao locutor *n*, e ε_n o resíduo. De toda a variabilidade associada ao contexto, uma parte pode ser modelada pelas variáveis de contexto indicadas no Quadro 1. Dessa forma, a variabilidade de contexto X_C passa a ter uma parcela modelada X_{CM} e uma parcela não modelada X_{CN}. Expandindo a Equação 1 e agrupando as variabilidades não modeladas pelas variáveis de contexto do Quadro 1, tem-se que

$$Y_n \approx X_{CM|n} + \left(X_{L|n} + X_{CN|n} + \varepsilon_n\right) \approx X_{CM|n} + \varepsilon_{L,CN|n}, \qquad \text{(Equação 2)}$$

onde $\varepsilon_{L,CN|n}$ é o resíduo do modelo quando modelado apenas pelas variáveis do Quadro 1. Esse resíduo incorpora a variabilidade de contexto não modelada pelas variáveis do Quadro 1 e a variabilidade do locutor.

Para cada uma das medidas acústicas indicadas no Quadro 2, foi ajustado um modelo linear generalizado do tipo

$$\begin{aligned} \overline{Y}_n &= \beta_0 + BX_{CM|n} \\ Y_n &\sim N\left(\overline{Y}_n, \varepsilon_{L,NC|n}\right). \end{aligned} \qquad \text{(Equação 3)}$$

O modelo assume que o valor médio Y_n de uma medida acústica depende de um termo independente β_0 mais a combinação linear das $\boldsymbol{X}$ variáveis de contexto ponderadas pelo vetor $\mathbf{B}$. Os valores das medidas Y_n são oriundas de um distribuição normal com média Y_n e desvio padrão $\varepsilon_{L,CN|n}$.

Dentro das medições e etiquetamentos realizados, notou-se que as variáveis que representavam o som anterior e o som posterior à unidade amostral medida apresentavam conjuntos categóricos muito diversos. Registraram-se 33 categorias para anterior e 27 para posterior. Assim como a ocorrência das vogais não é homogênea (vide Figura 1), a ocorrência das categorias nessas variáveis apresentou heterogeneidade. Outro fato é que nem todas as categorias ocorriam em todos os locutores. Esse fato dificultava estabelecer um conjunto de treinamento e um de testes para o ajuste do GLM.

Para contornar cada um dos sons, foi transformado em cinco variáveis fictícias binárias (*dummy variable*). Relacionadas com os seguintes fatores:

1. Obstrução: que indica se o som é emitido livremente. Definiu-se o valor 1 para vogais e 0 para consoantes.
2. Vozeamento: Indica se o som é executado com a vibração das pregas vocais. Em todas as vogais e nas consoantes vozeadas, o valor é 1; nos sons não vozeados, é 0.
3. Abertura da boca: Indica se o som é produzido com a boca mais aberta.

Definiu-se valor 0 para as vogais altas e para as consoantes plosivas, e 1 para os demais sons.

4. Posição dos articuladores: Apresenta valor 1 para as consoantes articuladas na posição frontal da boca (i.e., lábios, dentes ou alvéolo) e para as vogais frontais. O valor 0 é atribuído em caso contrário.

5. Nasalidade: apresenta valor 1 para as vogais e consoantes nasais e 0 caso contrário.

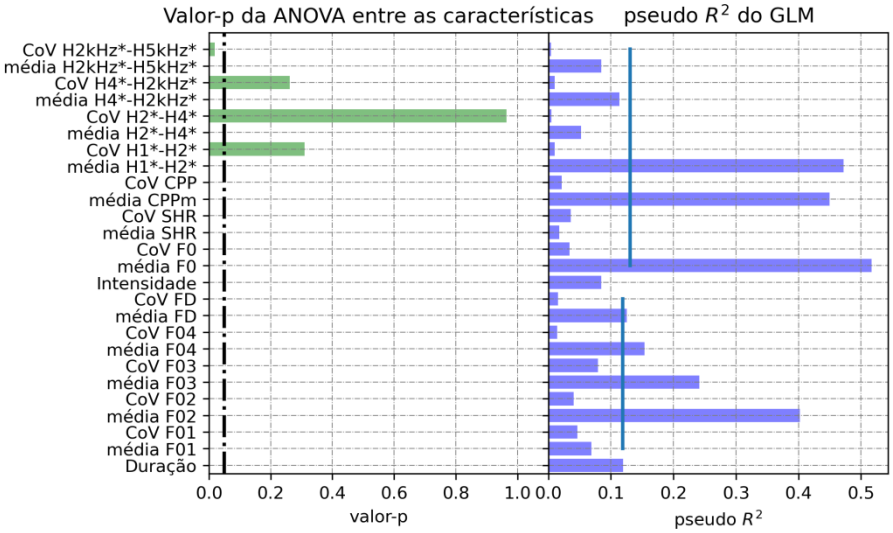
6. Em caso de ausência de som anterior ou precedente, todas as variáveis fictícias assumem valor igual a 0.

A análise das medidas acústicas no espaço observável indicou a presença de variabilidade intra e extrafalante. A análise de variância (ANOVA – *Analysis of Variance*) das medidas acústicas agrupadas pelos locutores indicou que separadamente apenas três tipos de medida são capazes de distinguir pelo menos o locutor. A Figura 4 apresenta, no gráfico da esquerda, o valor-*p* da análise de variância em que as medidas Cov H1*–H2*, Cov H4*–H2kHz* e Cov H1*–H2* apresentaram valor-*p* acima de 0,05. No gráfico à direita da Figura 4, as barras horizontais indicam os valores do pseudocoeficiente de determinação (pseudo R^2) relativo ao GLM de cada medição acústica, enquanto as linhas verticais marcam as médias relativas às categorias de medidas acústicas articulatórias e vocais.

Analisando o pseudo R^2 do modelo linear generalizado de cada medida acústica, nota-se uma média de 0,124, em que a maioria dos valores se encontra abaixo de 0,20. Isso indica que os modelos baseados em contexto foram capazes de explicar, na média, 12,4% da variância dos dados. Por outro lado, as médias de H1*–H2*, CPP, F_0 F_3 e F_2 foram capazes de explicar acima de 20% da variância.

Ao comparar o pseudo R^2 desses dois grupos, o teste de Kolmogorov-Smirnov falha em rejeitar que as distribuições são diferentes (valor-*p* de 0,31). O teste de Levene de igualdade de variância também falha em indicar a diferença de variância (valor-*p* de 0,59). O teste t de diferença entre as médias também falha (valor-*p* de 0,87). Os referidos testes indicam que, quando observados separadamente, os modelos dos grupos articulatório e vocal não apresentam diferença significativa do pseudo R^2.

Figura 4 – No gráfico à esquerda, o valor-p da análise de variância entre os locutores de acordo com cada característica. Notam-se apenas três características isoladas que podem distinguir pelo menos um locutor. À direita, o pseudocoeficiente de determinação do GLM com as linhas azuis verticais mostra a média das variáveis articulatórias e vocais.



Fonte: elaborado pelos autores.

3.2 Aplicação a Comparação de Locutores

Uma consequência da aplicação do GLM na modelagem da variabilidade das medidas acústicas é a informação residual. Como indicado nas equações 2 e 3, parte da variabilidade das variáveis de contexto X_{CM} são incorporadas no modelo, enquanto parte da variabilidade, incluindo do locutor, são incorporadas ao resíduo $\varepsilon_{L,CN|n}$. Partindo dessa premissa, espera-se que uma comparação de locutor no espaço dos resíduos apresente uma variabilidade menor que uma comparação de locutor no espaço das medidas acústicas. Visando testar esta hipótese, planejou-se um experimento que aplicou uma metodologia de CFL a partir dos três espaços de medidas acústicas: (1) o espaço das variáveis mensuráveis (doravante etiquetado como "**VA**"); (2) o espaço das componentes principais (doravante etiquetado como "**PCA**"); e (3) o espaço dos resíduos do GLM (doravante etiquetado como "**GLM-RES**"). Para o espaço dos resíduos do GLM, ainda foi realizada uma subdivisão separando as medidas articulatórias (doravante etiquetado como "**GLM-RES-ART**") e vocais (doravante etiquetado como "**GLM-RES-VOC**"). A Figura 5 apresenta um diagrama de blocos que indica as etapas do procedimento de obtenção dos resíduos do modelo GLM com destaque para a etapa comum de comparação de locutores.

A metodologia empregada para a comparação dos locutores é baseada nas etapas sequenciais:

1. normalização do espaço pela média e desvio padrão dos dados;
2. geração de duas subamostras, a de treinamento e a de testes obtidas por

bootstrap;

3. utilização da amostra de treinamento para o cálculo da distância euclidiana entre as subamostras, indicando as comparações realizadas entre mesmo locutor e locutores diferentes;

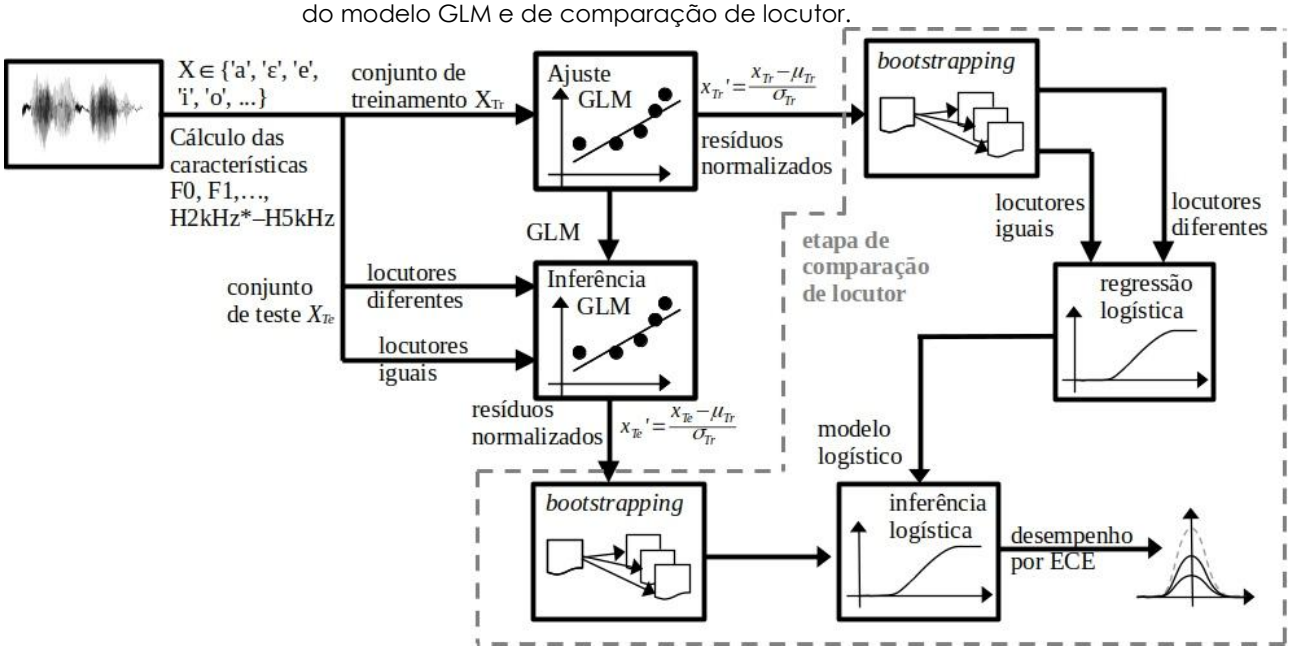
4. ajuste de um modelo de regressão logística com base nas duas classes de comparações, mesmo locutor e locutores diferentes, utilizando o conjunto de treinamento; e

5. validação do modelo, com o conjunto de teste, e cálculo das métricas de desempenho.

A normalização dos dados visa homogeneizar o espaço de comparação. Como as medidas acústicas são representadas em diferentes unidades e em escalas diferentes, a normalização homogeniza os valores para média nula e desvio padrão unitário. Essa normalização é uma das etapas da transformação das medidas acústicas para o espaço de componentes principais (PCA).

A geração das duas subamostras, uma de treinamento e outra de teste, visa dividir as etapas e homogeneizar o número de unidades de comparação entre os locutores. Observando a Tabela 2, nota-se que o número de unidades amostrais por locutor varia entre 143 e 512. Visando capturar a informação de cada locutor e homogeneizar a rotina de comparação, cada locutor foi sintetizado em 20 amostras, sendo 14 de treinamento e 6 de teste. Cada amostra é a média aritmética de uma subamostra aleatória de 20% de todas as medidas acústicas de cada locutor. Esse procedimento de subamostragem, denominado *bootstrap*, tende a preservar a média e a dispersão da amostra original, além de homogeneizar as amostras por locutor e reduzir o esforço computacional.

Figura 5 – Diagrama de blocos indicando as etapas do procedimento de obtenção dos resíduos do modelo GLM e de comparação de locutor.



Fonte: elaborado pelos autores.

Com a amostra de treinamento, composta por 252 (14x18) vetores, são calculadas as 31.626 distâncias euclidianas com a indicação das 1.638 realizadas entre locutores iguais e das 29.988 realizadas entre locutores diferentes. Em geral, os experimentos de comparação em grupos com muitos locutores tendem a ter uma prevalência reduzida, que, neste caso, é de 5,2%.

Com base nos valores das distâncias euclidianas, ajustou-se um modelo de regressão logística considerando os dois grupos – mesmo locutor e locutores diferentes. Do modelo de regressão logística, extraiu-se o limiar de decisão entre os grupos e a taxa de mesmo erro (EER – *equal error rate*). A distribuição dos valores de razão de verossimilhança dos grupos possibilitou o cálculo do custo do logaritmo da razão de verossimilhança C_{LLR} e da curva de entropia empírica cruzada (ECE – *empirical cross entropy*) para cada um dos três espaços de medidas acústicas e as divisões entre articulatória vocal por resíduos do GLM.

A partir do conjunto de testes, utilizando a validação cruzada, calculou-se a acurácia de comparação juntamente com as taxas de falso positivo (TFP, associado ao erro do tipo I) e a taxa de falso negativo (TFN, associado ao erro do tipo II). O resultado da etapa de treinamento e de testes é apresentado na Tabela 4. Na etapa de treinamento, a comparação realizada no espaço dos resíduos do GLM (GLM-RES) apresentou o melhor desempenho seguido pelas subdivisões pelas medidas vocais (GLM-RES-VOC) e articulatórias (GLM-RES-ART). Na etapa de testes, esses espaços de medidas acústicas também apresentaram o melhor desempenho em acurácia (99,1%) e na TFP (0,6%). Por outro lado, apresentaram a pior performance na taxa de falso negativo.

Tabela 4 – Índices de desempenho das etapas de treinamento e de testes. Na etapa de testes, os valores entre parênteses indicam o intervalo de confiança da medida com confiabilidade de 95%. Destacaram-se, em negrito, os resultados de melhor desempenho.

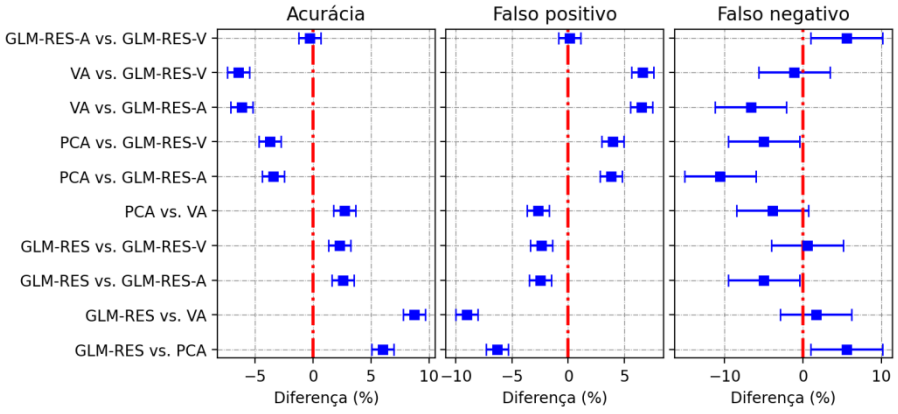
Espaço da medida acústica	Treinamento		Teste (intervalo de confiança)		
	EER (%)	C_{LLR} (np)	Acurácia (%)	TFP (%)	TFN (%)
VA	8,9	0,271	89,8 (88,9; 90,6)	10,2 (9,3; 11,1)	10,0 (6,8; 13,2)
PCA	4,8	0,175	93,2 (92,7;93,8)	6,8 (6,2; 7,4)	**5,8 (3,4; 8,3)**
GLM-RES	**0,1**	**0,007**	**99,1 (98,9;99,3)**	**0,6 (0,4; 0,8)**	11,7 (9,6; 13,7)
GLM-RES-ART	3,0	0,078	96,5 (96,1; 96,9)	3,0 (2,6; 3,4)	21,1 (18,5; 23,7)
GLM-RES-VOC	2,2	0,068	96,9 (96,5; 97,2)	2,8 (2,5; 3,2)	12,2 (10,2; 14,2)

Fonte: Elaborado pelos autores.

Ao realizar uma análise de variância pelo teste da diferença honestamente significativa (HDS - *honestly significant difference*) de Tukey, foram obtidos os seguintes resultados: (1) que o desempenho da comparação de locutores no espaço GLM-RES apresentou uma acurácia significativamente superior; e (2) uma taxa de falso positivo significativamente inferior à comparação realizada nos demais espaços. A diferença na taxa de falso negativo foi significativa apenas em relação ao PCA. A Figura 6 apresenta a comparação entre os resultados da Tabela 4 para a etapa de testes.

Outro resultado é que a acurácia e a taxa de falso positivo são estatisticamente equivalentes nos espaços GLM-RES-ART e GLM-RES-VOC, com diferença significativa na taxa de falso positivo. Na curva da entropia cruzada empírica (vide Figura 7a), nota-se que a comparação de locutores no espaço GLM-RES apresentou um desempenho superior às medidas nos demais espaços.

Figura 6 – Resultado da análise de variância da acurácia e as taxas de falso positivo e falso negativo para a comparação dos locutores nos diferentes espaços de características.



Fonte: elaborado pelos autores.

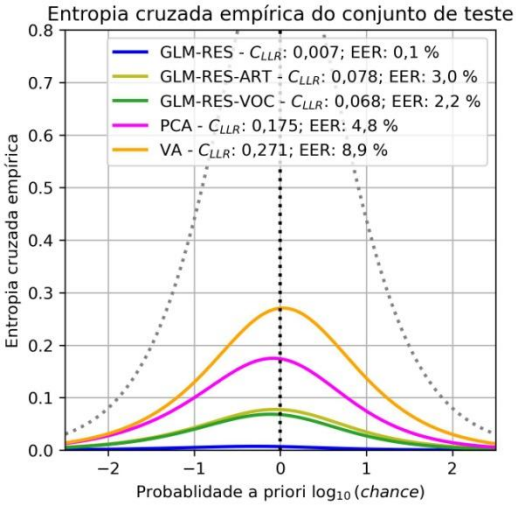
Para avaliar a variabilidade inter e extrafalante, utilizou-se a medida de distância euclidiana. Para cada grupo de amostras de um locutor Z_n

(subdivididos nas amostras de treinamento e teste), a variabilidade intrafalante foi calculada como a média das distâncias euclidianas de cada subamostra para o centroide do locutor. A variabilidade extrafalante foi calculada como a média das distâncias dos centroides de cada locutor para o centroide da subamostra (todos os locutores). Os valores foram normalizados pela maior distância extrafalante.

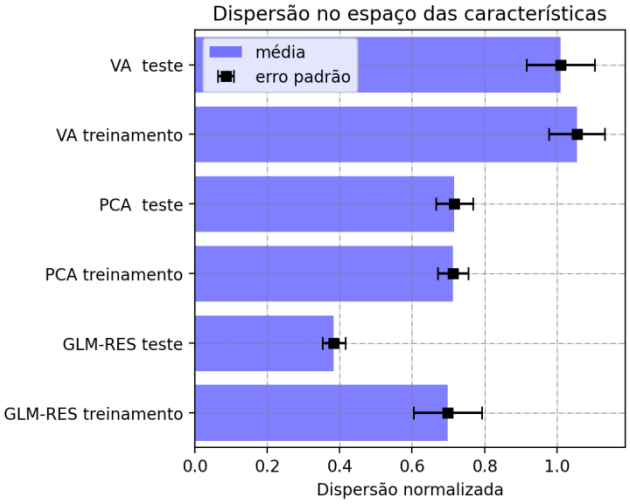
Uma separação dos locutores eficiente apresenta a variabilidade intrafalante minimizada e uma variabilidade extrafalante maximizada. Definiu-se como razão de dispersão a divisão entre a medida de variabilidade intrafalante pela variabilidade extrafalante. A Figura 7b indica nas barras horizontais o valor da média da razão de dispersão juntamente como o intervalo do erro padrão. O resultado mostra que o espaço GLM-RES apresentou a menor razão de viabilidade.

Uma forma de visualizar a dispersão é utilizando o escalonamento multidimensional (MDS - *multidimensional scaling*). O MDS é uma técnica que permite projetar um espaço em uma dimensão menor preservando a proporção das distâncias entre os pontos. A Figura 8 apresenta a dispersão das amostras de teste no espaço MDS referente a cada tipo de medida acústica. Na imagem, nota-se que, no espaço PCA (gráfico inferior à esquerda), ocorre uma separação dos grupos do sexo feminino e masculino e que ocorre uma menor sobreposição das amostras em relação ao espaço das variáveis acústicas mensuráveis (gráfico superior à direita). No espaço dos resíduos do GLM (gráfico superior à direita), nota-se que as amostras de cada locutor estão com uma dispersão menor (menor variabilidade intrafalante), que não ocorre um agrupamento pelo sexo do falante e que a dispersão extrafalante é maior.

Figura 7 – Desempenho das etapas de treinamento e teste da comparação de locutor. À esquerda, as curvas de entropia cruzada empírica, e à direita a razão de dispersão média e o erro padrão para cada subamostra.



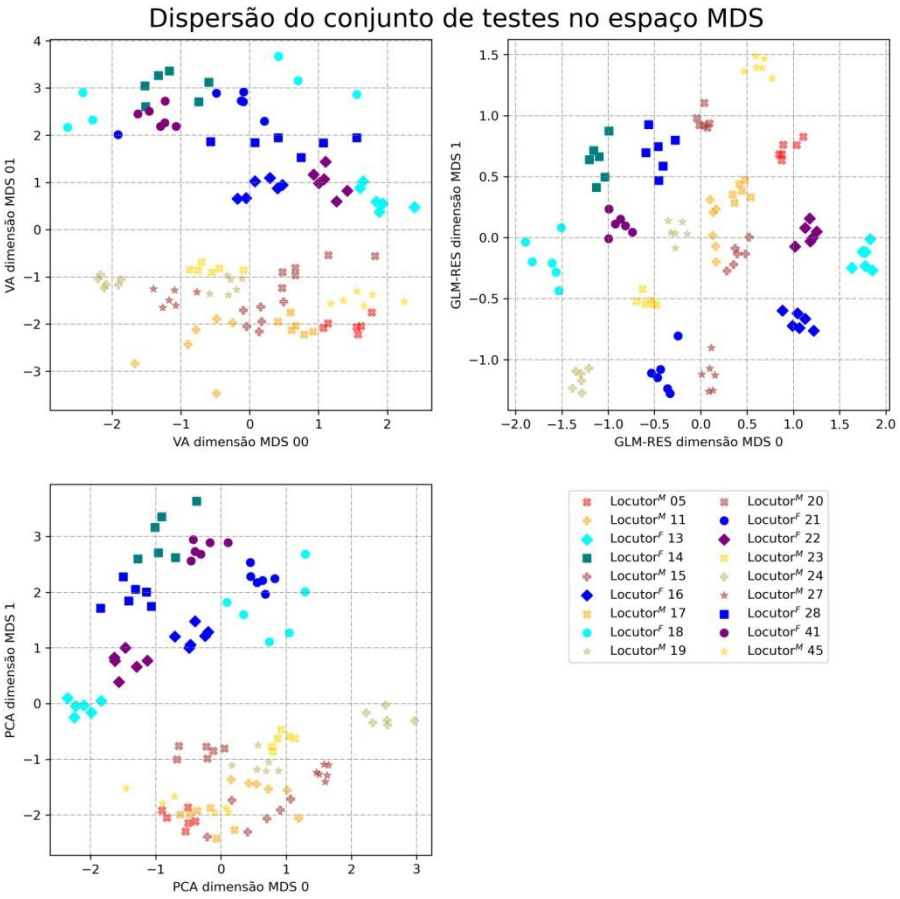
(a) Entropia cruzada empírica do resultado da comparação de locutores nos três espaços de variáveis.



(b) Razão de dispersão das subamostras de treinamento e de teste.

Fonte: elaborado pelos autores.

Figura 8 – Gráfico de dispersão das amostras dos locutores em cada espaço de medidas acústicas projetadas em duas dimensões por escalonamento multidimensional (MDS).



Fonte: elaborado pelos autores.

4 DISCUSSÃO

Primeiramente, é importante pontuar que o experimento apresenta um número reduzido de locutores (18), fato que limita a generalização de parte dos resultados. Essa limitação deve-se ao fato de o etiquetamento das vogais e dos ditongos ser realizado por trabalho manual. O processo como um todo exige atenção do profissional que executa e uma revisão atenta. As 5.615 vogais e ditongos foram selecionados entre várias etiquetagens que apresentavam erros e inconsistências. Outra limitação é de um número reduzido de variáveis de contexto presentes no estudo (vide Quadro 1) e a presença de apenas um dialeto e nenhuma variação do estado do locutor como emoção ou patologia.

Dadas as limitações acima, a modelagem por contexto apresentou na média um pseudo R^2 de 0,124. Apesar de, na média, o valor ser baixo, algumas medidas acústicas apresentaram pseudo R^2 acima de 0,4. Em relação à variabilidade intra e extralocutor, o experimento mostrou que a aplicação do GLM é capaz de reduzir a variabilidade intralocutor e explica que fatores como a tonicidade, posição, e os sons adjacentes influenciam nos valores obtidos nas medidas acústicas. O sexo do locutor, apesar de não ser um fator puramente contextual, também contribui para explicar parte da variabilidade.

Em relação à variabilidade extralocutor, a aplicação do GLM também se mostrou eficiente. Os resultados mostram uma redução da razão de dispersão na ordem de 63% em relação às variáveis mensuráveis e de 46% em relação às variáveis no espaço das componentes principais.

Sobre a comparação de locutores, os valores obtidos de EER e de C_{LLR} são da ordem de grandeza de resultados obtidos no estado da arte, como por Sztahó e Fejes (2023), que utilizaram características de gargalo obtidas de redes neurais profundas ou de Ishihara (2021), que utiliza grupos de palavras. Entretanto, esse desempenho precisa de mais desenvolvimento para ser aplicado em situações forenses reais.

Outro ponto que os autores gastariam de citar foi o valor da taxa de falso positivo do espaço GLM-RES, apesar do índice apresentar o pior desempenho 11,7% (101% acima do melhor resultado). Entretanto, o falso negativo – que é deixar de associar dois locutores quando as vozes analisadas são oriundas do mesmo locutor – pode ser menos prejudicial devido ao princípio do *in dubio pro reu*. Em relação à origem das medidas, articulatório ou vocais, os subgrupos não apresentaram diferenças significativas, sendo mais evidente o efeito da combinação.

5 CONCLUSÃO

Em relação à modelagem da variabilidade relacionada ao falante, o experimento mostrou que, em média, 12,4% da variabilidade está relacionada ao contexto modelado e que tanto as estruturas articulatórias quanto vocais apresentam contribuição que não são significativamente diferentes. Em relação à classificação dos locutores, ambas as estruturas se mostraram muito semelhantes, sendo que as vocais superam as articulatórias em relação à taxa de falso negativo, quando dispostas em modelos isolados.

O experimento mostrou-se eficaz na tarefa de remover do sinal acústico parte da informação referente ao contexto fonológico e acentuar a informação da identidade do locutor. Porém, não pode ser extrapolado devido às limitações de amostra e de dialetos. Por outro lado, o experimento indicou que mais investigações necessitam ser realizadas.

Como propostas de continuidade, o presente trabalho busca desenvolver um método automatizado de etiquetamento das vogais e de estabelecimento do contexto. Para a expansão do experimento, planeja-se variar o dialeto e incluir variáveis de contexto como prosódia e cadência. Em relação às medidas acústicas, planeja-se expandir a lista de medidas com, por exemplo, medidas baseadas em *ceptrum,* além de incluir variações temporais como inclinação e concavidade.

REFERÊNCIAS

CAMPBELL, J. P.; SHEN, W.; CAMPBELL, W. M.; SCHWARTZ, R., BONASTRE, J. F.; MATROUF D. Forensic speaker recognition. **IEEE Signal Processing Magazine**, v. 26, n. 2, p. 95-103, 2009.

CHAMBERS, J. K. **Sociolinguistic theory:** Linguistic variation and its social significance. Wiley, 1995.

DODDINGTON, G. Speaker recognition based on idiolectal differences between speakers. In: **Seventh European Conference on Speech Communication and Technology**. 2001.

DRYGAJLO, A. Forensic evidence of voice. **Encyclopedia of biometrics**, p. 1388-1395, 2009.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**, 2nd edition. New York, USA: John Wiley&Sons, v. 35, 2001.

FANT, G. **Acoustic theory of speech production:** with calculations based on X-ray studies of Russian articulations. [S.l.]: Walter de Gruyter, 1971. v. 2.

FITCH, W. T. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. **The Journal of the Acoustical Society of America**, v. 102, n. 2, p. 1213-1222, 1997.

FLANAGAN, J. L. **Speech analysis synthesis and perception**. Springer Science & Business Media, 2013.

FURUI, S. **Digital speech processing: synthesis, and recognition**. CRC Press, 2018.

GARELLEK, M. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. **Journal of Phonetics**, v. 94, p. 101155, 2022.

GFRÖRER, S. G. Auditory-instrumental forensic speaker recognition. In: **INTERSPEECH**. 2003. p. 705-708.

GONZALEZ-RODRIGUEZ J, ROSE P, RAMOS D, TOLEDANO DT, ORTEGA-GARCIA J. Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 15, n. 7, p. 2104-2115, 2007.

HANSON, H. M.; CHUANG, E. S. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. T**he Journal of the Acoustical Society of America**, v. 106, n. 2, p. 1064-1077, 1999.

HILLENBRAND, J.; CLEVELAND, R. A.; ERICKSON, R. L. Acoustic correlates of breathy vocal quality. **Journal of Speech, Language, and Hearing Research**, v. 37, n. 4, p. 769-778, 1994.

ISELI, M.; ALWAN, A. An improved correction formula for the estimation of

harmonic magnitudes and its application to open quotient estimation. In: **IEEE international conference on acoustics, speech, and signal processing**. IEEE, 2004. p. I-669.

ISHIHARA, S. Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. **Forensic Science International**, v. 327, p. 110980, 2021.
KABIR, M. M.; MRIDHA, M. F.; SHIN, J.; JAHAN, I.; OHI, A.Q. A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. **IEEE Access**, v. 9, p. 79236-79263, 2021.

KERSTA, L. G. Voiceprint identification, **Nature**, vol. 196, no. 4861, pp. 1253-1257, 1962.

KILBOURN-CERON, O.; GOLDRICK M. Variable pronunciations reveal dynamic intra-speaker variation in speech planning. en. In: **Psychonomic Bulletin & Review** 28.4, pp. 1365–1380, 2021.

KREIMAN, j.; SIDTIS, D. **Foundations of voice studies: an interdisciplinary approach to voice production and perception**. Malden, MA: Wiley-Blackwell, 2011.

KREIMAN, J.; PARK, S. J.; KEATING, P. A.; ALWAN, A. The relationship between acoustic and perceived intraspeaker variability in voice quality. In: **Sixteenth Annual Conference of the International Speech Communication Association**. 2015.

LABOV, W. **Sociolinguistic patterns**. University of Pennsylvania press, 1973.

LAVAN, N.; BURSTON, Luke F. K.; GARRIDO, L. How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. **British Journal of Psychology**, v. 110, n. 3, p. 576-593, 2019.

LEE, Y.; KEATING, P.; KREIMAN, J. Acoustic voice variation within and between speakers. **The Journal of the Acoustical Society of America**, v. 146, n. 3, p. 1568-1579, 2019.

MAHER, R. C. Audio forensic examination. **IEEE Signal Processing Magazine**, v. 26, n. 2, p. 84-94, 2009.

MCQUISTEN, K. A.; PEEK, A. S. Comparing artificial neural networks, general linear models and support vector machines in building predictive models for small interfering RNAs. ***PLoS One***, 4(10), e7522, 2009.

MORRISON, G. S. Forensic voice comparison and the paradigm shift. **Science & Justice**, v. 49, n. 4, p. 298-308, 2009.

NETO, A. F.; SILVA, A. P.; YEHIA, H. C. Corpus CEFALA-1: base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia/corpus CEFALA-1: audiovisual database of speakers for biometric, phonetic and phonology studies. **Revista de Estudos da Linguagem**, v. 27, n. 1, p. 191-212, 2019.

SAKS, M. J.; KOEHLER, J. J. The individualization fallacy in forensic science evidence. **Vand. L. Rev.**, v. 61, p. 199, 2008.

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

SILVA, A. P.; VIEIRA, M. N.; BARBOSA, A. V. Avaliação de descritores acústicos em simulação de condições forenses de verificação de locutor. **Revista Brasileira de Crimianlística**, v. 8, n. 2, p. 22-35, 2019.

SILVA, A. P. **Intervalo de evidência e pareamento fuzzy utilizando relação sinal ruído aplicados à comparação forense de locutores**. 2020. Tese de Doutorado. Universidade de Federal de Minas Gerais.

SUN, X. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In: **2002 IEEE international conference on acoustics, speech, and signal processing**. IEEE, 2002. p. I-333-I-336.

SZTAHÓ, D.; FEJES, A. Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings. **Journal of forensic sciences**, v. 68, n. 3, p. 871-883, 2023.

WICKHAM, H. *Tidy Data.* ***Journal of Statistical Software***. Vol. 59 (10), 2014. doi:10.18637/jss.v059.i10.

WÜTHRICH, M. V. From generalized linear models to neural networks, and back. ***SSRN,*** *Manuscript ID, 3491790, 2019.*