

Modelagem estatística da variabilidade de inter e intrafalantes em fala contínua

Maria Mandes Cantoni & Adelino Pinheiro Silva

Núcleo de Linguística Computacional (ADA) / FALE/ UFMG
Ciência e Inteligência de Dados para Gestão e Segurança Pública (CIDaGESP) / ACADEPOL/ PCMG

6 de setembro de 2024



Sumário

- 1 Sumário
- 2 Introdução
- 3 Materiais e métodos
 - Análise de componentes principais
- 4 Metodologia de modelagem e resultados
 - Descrição Procedimental
 - Aplicação a comparação de locutores
- 5 Discussão
- 6 Considerações finais
- 7 Encerramento

Assuntos

- 1 Sumário
- 2 **Introdução**
- 3 Materiais e métodos
 - Análise de componentes principais
- 4 Metodologia de modelagem e resultados
 - Descrição Procedimental
 - Aplicação a comparação de locutores
- 5 Discussão
- 6 Considerações finais
- 7 Encerramento

Informação no sinal de fala

- Codificação;
 - identidade de grupo [Lab73];
 - identidade do falante [D⁺01, Ish21]
 - condições do falante;
 - contexto fonológico...
-
- Fonatórios (pregas vocais); e
 - Articulatórios (trato vocal). [Fan71, Fla13].

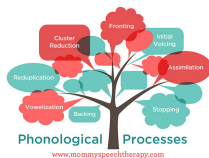


Imagem: <https://www.umsabadoqualquer.com/quem-realmente-deveria-estar-morto/>

Variabilidade intra e extrafalante

Intrafalante presente na fala de um mesmo locutor, devido as diferenças de como os movimentos da fala são articulados por ele;

Extrafalante: observada entre falantes distintos devido a diferentes tratos vocais e habilidades motoras [KCG21].



Imagens: <https://mommyspeechtherapy.com/?p=2158> e <https://aruno14.medium.com/speaker-recognition-using-tensorflow-1f007c2a7702>

Objetivo, perguntas e hipótese

Objetivo: modelar a variabilidade relacionada ao falante, consideram-se os papéis das estruturas articulatórias e vocais a partir de fala contínua.

Perguntas:

- 1 Quanto a 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?

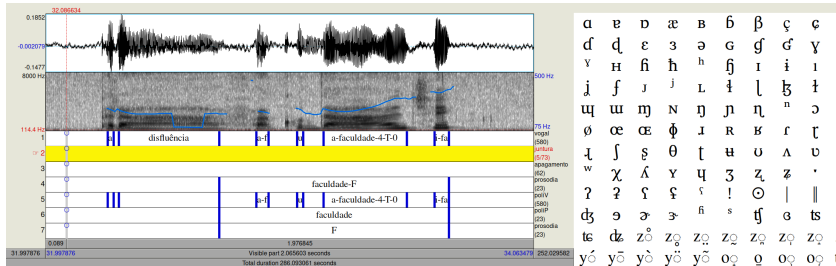
Hipótese: Um modelo linear generalizado pode remover informação do contexto fonológico e com isso acentuar a identidade do locutor.

Restrição: Modelagem estatística no espaço mensurável para ter mais controle e transparência em contraponto a RNA [MP09, Wüt19].

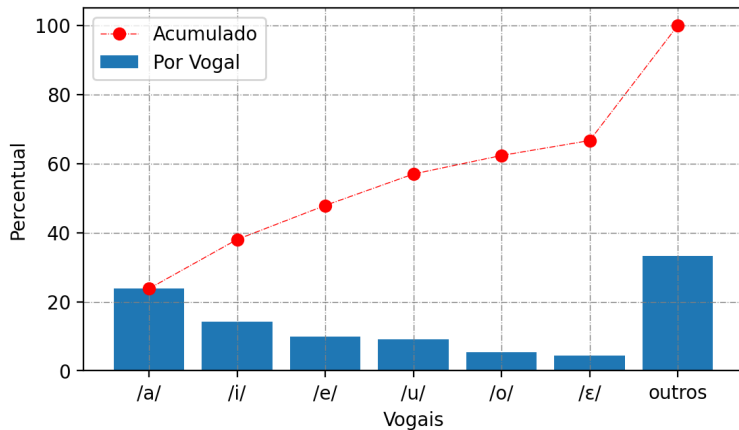
Assuntos

- 1 Sumário
- 2 Introdução
- 3 **Materiais e métodos**
 - **Análise de componentes principais**
- 4 Metodologia de modelagem e resultados
 - Descrição Procedimental
 - Aplicação a comparação de locutores
- 5 Discussão
- 6 Considerações finais
- 7 Encerramento

- 18 locutores (8 F e 10 M) do corpus CEFALA 1 [NSY19];
- 5615 unidades distribuídas em 64 categorias, sendo 12 vogais e 52 ditongos;
- entre 143 e 512 unidades por locutor;
- rotulagem manual em TextGrid no *praat* com revisão;
- processamento em *python 3.11* com pacotes *scipy*, *scikit-learn*, *pandas* e *matplotlib*.

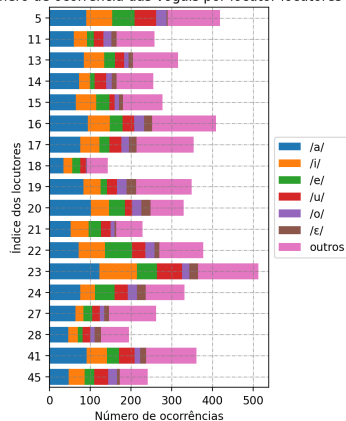


Distribuição das unidades amostrais



(a) Percentual de ocorrência das unidades amostrais.

Número de ocorrência das vogais por locutor locutores



(b) Ocorrência das unidades amostrais por locutor.

Variáveis de contexto

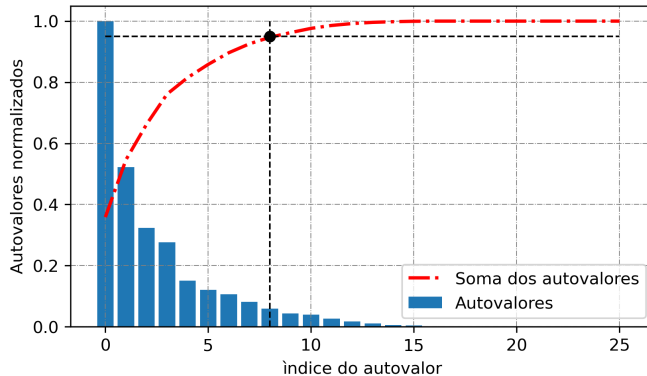
| Variável | Descrição |
|----------------------|--|
| Sexo do falante | Feminino ou masculino. |
| Tonicidade da sílaba | Posição relativa da sílaba em relação a tônica, os valores variam entre tônica, pré-tônica ou pós-tônica. |
| Posição na palavra | Sílaba da palavra, a partir do começo, a vogal se encontra. Os valores na amostra variam entre 0 e 6. |
| Ditongação | Indica se a unidade amostral é uma vogal com valor 0 ou ditongo com valor 1. |
| Fechamento | Sílaba onde se encontra a vogal (ou ditongo) é fechada por consoante. |
| Som anterior | Som fonológico que ocorre antes da vogal, valor nulo (final de palavra), vogal (e.g., /a/, /ε/, /ĩ/, ...) ou consoante (e.g., /p/, /tʃ/, /n/, ...). |
| Som posterior | Som fonológico que ocorre depois da vogal, valor nulo (final de palavra), vogal (e.g., /a/, /ε/, /ĩ/, ...) ou consoante (e.g., /p/, /tʃ/, /n/, ...). |

Medidas acústicas

Tabela: Lista de medidas conforme análise de [LKK19].

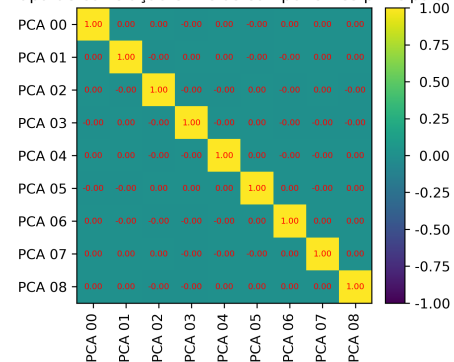
| Tipo de medida acústica | Tendência central | Variabilidade | Categoria |
|--|---|---|------------------|
| 1 – Intensidade e duração | Intensidade, Duração | – | Não classificada |
| 2 - Frequência dos formantes | média F_1 , média F_2 , média F_3 , média F_4 e média F_D | Cov F_1 , Cov F_2 , Cov F_3 , Cov F_4 e Cov F_D | 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 |

Distribuição das componentes principais



(a) Autovalores normalizados da matriz de correlação.

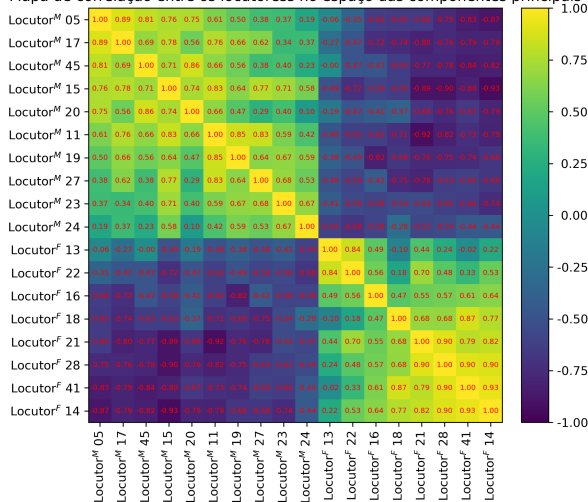
Mapa de correlação entre as componentes principais



(b) Mapa de correlação entre as nove primeiras CP.

Correlação no espaço de componentes principais

Mapa de correlação entre os locutores no espaço das componentes principais



Assuntos

- 1 Sumário
- 2 Introdução
- 3 Materiais e métodos
 - Análise de componentes principais
- 4 Metodologia de modelagem e resultados**
 - Descrição Procedimental
 - Aplicação a comparação de locutores
- 5 Discussão
- 6 Considerações finais
- 7 Encerramento

Equacionamento

Modelos linear generalizado (GLM - *generalized linear model*)

$$Y_n \approx X_{C|n} + X_{L|n} + \epsilon_n \quad (1)$$

$$Y_n \approx X_{CM|n} + X_{CN|n} + X_{L|n} + \epsilon_n \approx X_{CM|n} + \epsilon_{L,CN|n} \quad (2)$$

$$\begin{aligned} \bar{Y}_n &= \beta_0 + \mathbf{B} \cdot \mathbf{X}_{CM|n} \\ Y_n &= \mathcal{N}(\bar{Y}_n, \epsilon_{L,CN|n}) \end{aligned} \quad (3)$$

Criação de variáveis fictícias

Número de Categorias dos sons anteriores e posteriores

Homogeneizar a ocorrência em todos os locutores

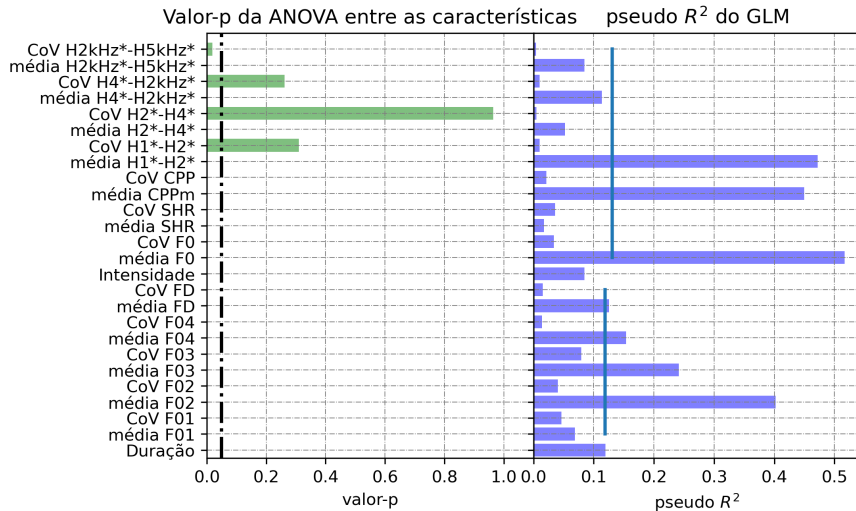
- 1 Obstrução: valor 1 para vogais e 0 para consoantes.
- 2 Vozeamento: em todas vogais e nas consoantes vozeadas é 1 e 0 nos sons não vozeados.
- 3 Abertura da boca: Valor 0 para as vogais altas e para as consoantes plosivas e 1 caso contrário.
- 4 Posição dos articuladores: Apresenta valor 1 para as consoantes articuladas na posição frontal da boca (i.e., lábios, dentes ou alvéolo) e para as vogais frontais.
- 5 Nasalidade: apresenta valor 1 para as vogais e consoantes nasais.

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

Ajuste do modelo

Conjunto de treinamento (70%) e de teste (30%).

Homogenização de 20 vetores por locutor (14/6) por *bootstrap*.



Dispersão no espaço MDS

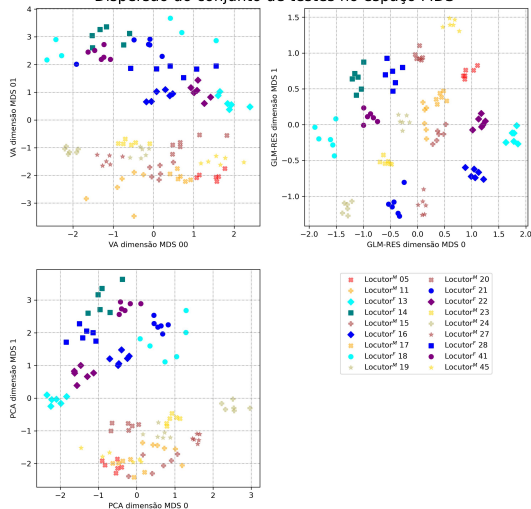
Pseudo R^2 vocais vs. articulatórias

Kolmogorov-Smirnov falha em rejeitar que as distribuições são diferentes (valor-p de 0,31).

Levene falha em indicar a diferença de variância (valor-p de 0,59).

O **teste t** de diferença entre as médias falha (valor-p de 0,87).

Dispersão do conjunto de testes no espaço MDS



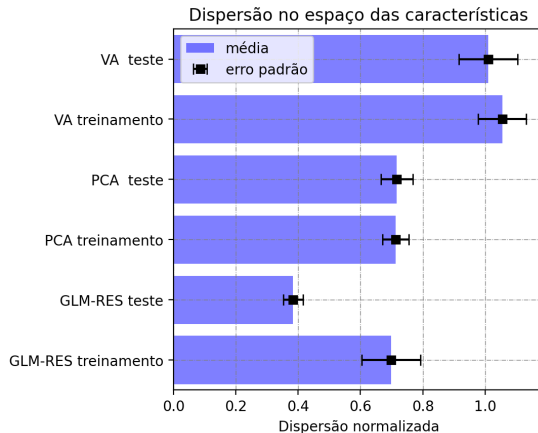
Razão de dispersão

Intralocutor: D_w distâncias de cada conjunto de vetores ao centroide.

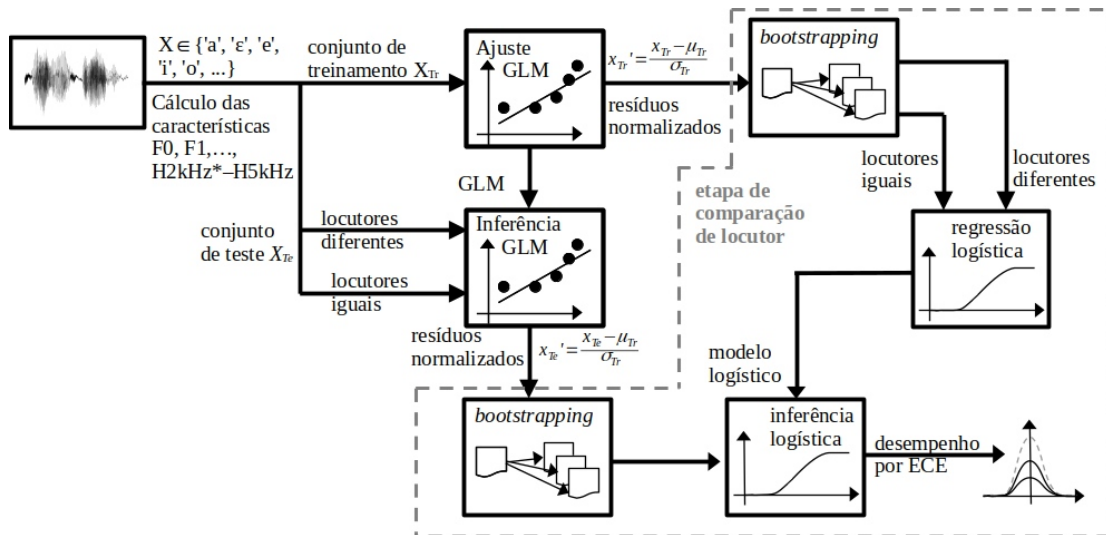
Interlocutor: D_b distância do centro da amostra a cada centroide de locutor normalizada pela distância do centroide mais distante*.

$$RD = \frac{D_w}{D_b}$$

* a normalização foi para fazer o ajuste da dimensionalidade.



Fluxograma das etapas de modelagem e comparação



Etapas da comparação de locutor

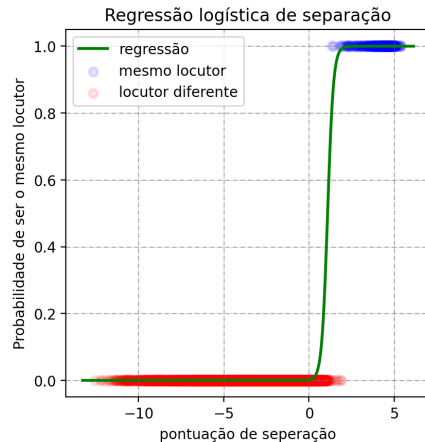
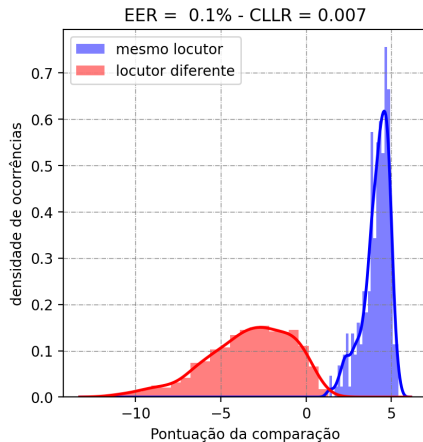
A metodologia empregada para a comparação dos locutores é baseada nas etapas sequenciais:

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

Regressão logística

$$LR = \frac{P(ML)}{P(LD)}$$

$$LLR = \log_2 \left(\frac{P(ML)}{P(LD)} \right)$$

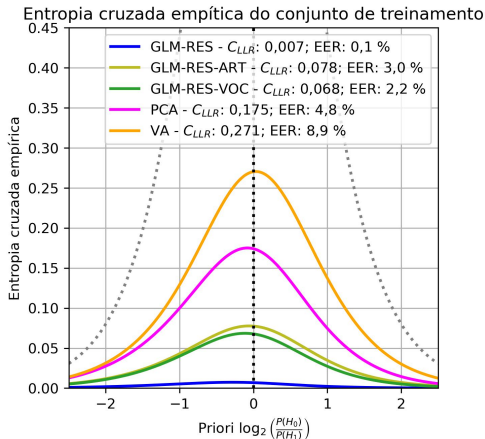


Desempenho de ECE

Pontilhada cinza: Classificador aleatório (LR = 1).

Quanto mais “baixa” menor o custo do erro na classificação.

$$\begin{cases} H_0 : & \text{Amostras locutores diferentes,} \\ H_1 : & \text{amostras do mesmo locutor.} \end{cases}$$



$$C_{LLR} = \frac{1}{2} \left(\frac{1}{N_{ML}} \sum \log_2 \left(1 + \frac{1}{LR_{MO}} \right) + \frac{1}{N_{LD}} \sum \log_2 (1 + LR_{LD}) \right).$$

Resultados da comparação de locutor

Taxa de mesmo erro (EER - *equal error rate*)

C_{LLR} : Custo do logaritmo da razão de verossimilhança

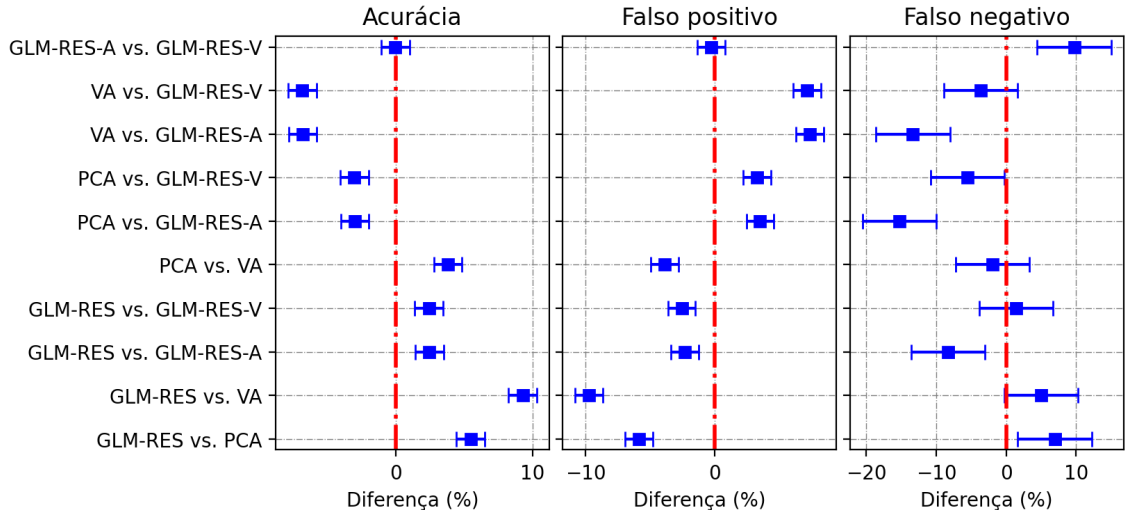
Entropia cruzada empírica (ECE - *empirical cross entropy*)

TFP: Taxa de falso positivo (erro tipo I).

TFN: Taxa de falso negativo (erro tipo II).

| Espaço da medida acústica | Treinamento | | Teste (intervalo de confiança) | | |
|---------------------------|-------------|------------------|--------------------------------|-----------------------|-----------------------|
| | EER (%) | C_{LLR} (bits) | 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 | 0,078 | 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) |

Análise de variância etapa de testes



Assuntos

- 1 Sumário
- 2 Introdução
- 3 Materiais e métodos
 - Análise de componentes principais
- 4 Metodologia de modelagem e resultados
 - Descrição Procedimental
 - Aplicação a comparação de locutores
- 5 Discussão**
- 6 Considerações finais
- 7 Encerramento

Principais achados

Resultados:

- O modelo GLM apresentou um pseudo R² de 0,124.
- Redução da razão de dispersão: 63% variáveis mensuráveis e de 46% PCA.
- Comparação de locutores com desempenho próximo ao estado da arte [SF23, Ish21].
- Pior falso positivo em 11,7% (101% acima do melhor resultado), aumenta o *in dubio pro reu*.

Ainda pode-se avaliar:

- Influência de cada variável de contexto;
- Influência das variáveis fictícias;

Assuntos

- 1 Sumário
- 2 Introdução
- 3 Materiais e métodos
 - Análise de componentes principais
- 4 Metodologia de modelagem e resultados
 - Descrição Procedimental
 - Aplicação a comparação de locutores
- 5 Discussão
- 6 Considerações finais**
- 7 Encerramento

Considerações finais

Principais pontos:

- No experimento específico foi eficiente em remover parte da variabilidade (++ hipótese);
- Foi eficiente na comparação dos locutores;
- Medidas vocais e articulatórias não apresentaram diferença significativa;
- Limitada na amostra de locutores e dialetos;

Continuidade:

- etiquetamento automatizado;
- expansão de dialetos;
- expandir contexto com cadência, prosódia, avaliação de emoção, etc...;
- expandir lista de medidas acústicas (e.g. *cepstrun*).

Assuntos

- 1 Sumário
- 2 Introdução
- 3 Materiais e métodos
 - Análise de componentes principais
- 4 Metodologia de modelagem e resultados
 - Descrição Procedimental
 - Aplicação a comparação de locutores
- 5 Discussão
- 6 Considerações finais
- 7 Encerramento**

Agradecimentos

Fim!

Contatos:

- e-mail: adelinocpp@gmail.com ou adelino.pinhoiro@policiacivil.mg.gov.br;
- Whatsapp (31) 98801-3605;
- Instituto de Criminologia - Academia de Polícia Civil de Minas Gerais. Rua Oscar Negrão de Lima nº 200, Nova Gameleira, Belo Horizonte-MG. Tel.: (31) 3314-5620.



Sobre este material

Esta obra está licenciada sob a licença *Creative Commons* CC BY-NC-SA 4.0






Favor fazer referência a este trabalho como:

Cantoni M. M., Silva, A. P., *Modelagem estatística da variabilidade de inter e intrafalantes em fala contínua*. Online: https://github.com/adelinocpp/silf_2024





```
@Misc{Cantoni2024,
title={Modelagem estatística da variabilidade de inter e intrafalantes em fala con
author={Maria Mendes Cantoni and Adelino Pinheiro Silva},
howPublished={\url{https://github.com/adelinocpp/silf_2024}},
year={2024},
note={Version 1.0; Creative Commons BY-NC-SA 4.0.},
}
```



Referências I

-  George R Doddington et al., *Speaker recognition based on idiolectal differences between speakers.*, Interspeech, 2001, pp. 2521–2524.
-  Gunnar Fant, *Acoustic theory of speech production: with calculations based on x-ray studies of russian articulations*, no. 2, Walter de Gruyter, 1971.
-  James L Flanagan, *Speech analysis synthesis and perception*, vol. 3, Springer Science & Business Media, 2013.
-  Shunichi Ishihara, *Score-based likelihood ratios for linguistic text evidence with a bag-of-words model*, Forensic Science International **327** (2021), 110980.
-  Oriana Kilbourn-Ceron and Matthew Goldrick, *Variable pronunciations reveal dynamic intra-speaker variation in speech planning*, Psychonomic Bulletin & Review **28** (2021), no. 4, 1365–1380.

Referências II

-  William Labov, *Sociolinguistic patterns*, no. 4, University of Pennsylvania press, 1973.
-  Yoonjeong Lee, Patricia Keating, and Jody Kreiman, *Acoustic voice variation within and between speakers*, The Journal of the Acoustical Society of America **146** (2019), no. 3, 1568–1579.
-  Kyle A McQuisten and Andrew S Peek, *Comparing artificial neural networks, general linear models and support vector machines in building predictive models for small interfering rnas*, PLoS One **4** (2009), no. 10, e7522.
-  Arlindo Follador Neto, Adelino Pinheiro Silva, and Hani Camille Yehia, *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 **27** (2019), no. 1, 191–212.

Referências III



Dávid Sztahó and Attila Fejes, *Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings*, Journal of forensic sciences **68** (2023), no. 3, 871–883.



Mario V Wüthrich, *From generalized linear models to neural networks, and back*, SSRN Manuscript ID **3491790** (2019).

Dúvidas



Imagem: <https://www.hevcon.com.br/duvidas-frequentes-relacionado-ao-novo-bem-e-as-alteracoes-das-mps/>.