Detecção da Fala Característica da Doença de Alzheimer Utilizando uma Abordagem Agnóstica de Idioma

Luan Dopke[1], João Paulo Aires[1], Juliana Onofre de Lira[2], Lilian Cristine Hübner[1], Dalvan Griebler [1]

[1] Escola Politécnica, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) Porto Alegre – RS – Brasil,

[2]Faculdade de Ciências e Tecnologias em Saúde, Universidade de Brasília (UnB) Brasília – DF – Brasil

luan.dopke@edu.pucrs.br, dalvan.griebler@pucrs.br

Abstract. *This study proposes a language-agnostic classification approach for detecting Alzheimer's disease based on speech features. Acoustic and textual features are extracted from English and Brazilian Portuguese data to train Machine Learning models. The results revealed that lexical indices and the ComParE feature set contribute to the model's generalization across languages.*

Resumo. *Este trabalho propõe uma abordagem de classificação agnóstica ao idioma para a detecção da doença de Alzheimer a partir de características da fala. Para isto são extraídas características acústicas e textuais de dados em inglês e português brasileiro para treinar modelos de Aprendizado de Máquina. Os resultados revelaram que os índices lexicais e o conjunto ComParE contribuem para a generalização do modelo entre os idiomas.*

1. Introdução

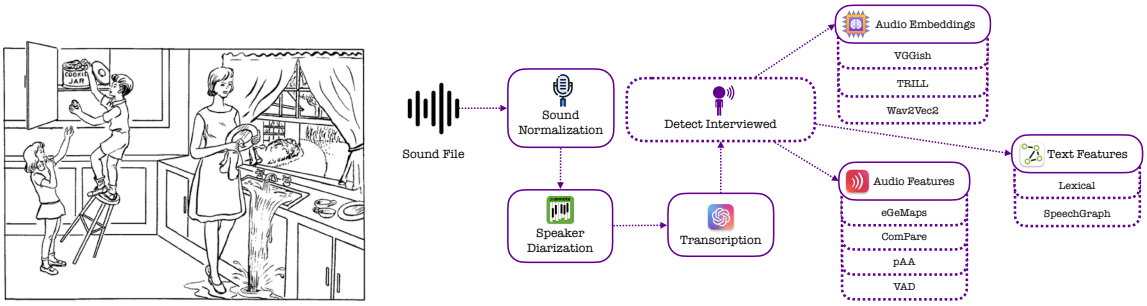
No estágio inicial da Doença de Alzheimer (DA), a linguagem é afetada, podendo causar redução do vocabulário e da fluência verbal. Considerando essas características, a construção de um classificador de fala característica de DA, capaz de identificar passivamente a condição, pode levar a um diagnóstico precoce da doença. Modelos capazes de distinguir indivíduos com DA de indivíduos sem a doença já estão presentes na literatura [Vigo et al. 2022]. Contudo, muitos desses modelos foram treinados com dados de um único idioma, limitando sua aplicabilidade apenas àquela língua específica. Também, muitos idiomas não possuem dados suficientes para treinar um modelo robusto. Um modelo que identifica padrões da fala da DA entre idiomas, capaz de distinguir estes de indivíduos sem a doença, pode preencher essa lacuna. Essa abordagem permite treinar um modelo em um idioma e transferir a capacidade preditiva para outro, solucionando o problema da escassez de dados em uma língua.

Considerando esses desafios, esta pesquisa tem como objetivo treinar e avaliar modelos de classificação para distinguir indivíduos com DA de indivíduos saudáveis a partir de características acústicas e textuais extraídas da fala que possam representar a DA. A contribuição deste trabalho é o desenvolvimento de um modelo que utiliza características de fala comuns ao inglês e ao português brasileiro para distinguir pacientes com DA de pessoas saudáveis, em uma abordagem que busca ser agnóstica ao idioma, focada na extração de características que transcendem as particularidades linguísticas de cada um.Assim, a questão de pesquisa é: Quais conjuntos de características acústicas e linguísticas podem ser extraídos de forma robusta para distinguir a fala de indivíduos com DA da fala de controles saudáveis nos idiomas inglês e português brasileiro?

2. Metodologia e Desenvolvimento

Para atingir nossos objetivos, foram utilizados dois conjuntos de dados em idiomas distintos — inglês e português brasileiro — contendo gravações da fala de pacientes com DA e controles saudáveis (CS). Para o idioma inglês, utilizou-se o conjunto de dados do desafio ADReSS-M [Luz et al. 2024], com 110 amostras de indivíduos com DA e 115 amostras de controles saudáveis. Enquanto que para o português brasileiro têm-se 20 gravações de pacientes com DA e 20 de controles saudáveis [1]. Ambos os conjuntos de dados contêm gravações de indivíduos descrevendo a imagem "Cookie Theft", Figura 1a.

Para treinar um modelo de classificação agnóstico ao idioma capaz de detectar DA, foi implementado um fluxo para a extração e avaliação de características: (1) Utilização a implementação do filtro EBU R128 do FFMPEG [2] para normalizar o áudio; (2) diarização para detecção dos locutores; (3) transcrição da fala; (4) separação da fala do pesquisadores e do entrevistado; (5) extração das características acústicas e linguísticas; (6) treinamento de modelos distintos para cada conjunto de características; (7) fusão tardia e, finalmente, (8) avaliação das combinações dos conjuntos de características.A Figura 1b exemplifica o processo de extração de características.



(a) Cookie Theft. **(b) Fluxo da extração dos conjuntos de características.**

Figura 1. A descrição da imagem (a) é utilizada como entrada no fluxo (b).

As gravações contêm a fala do entrevistado, mas também incluem os pesquisadores fornecendo orientações. Para manter apenas as falas do entrevistado, visando mitigar vieses, foi utilizado o WhisperX v3.3.1 [Bain et al. 2023] que usa o modelo Whisper[3] para transcrição e Pyannote.Audio[4] para a diarização, isto é, segmentar o áudio entre os períodos de fala e, dentro deles, atribuir um rótulo de identificação a cada locutor. Com base na diarização, determinou-se que o locutor que falou por mais tempo era o entrevistado, e o outro, o pesquisador.

A transcrição foi utilizada para capturar características textuais, através de métricas extraídas a partir de grafos de fala, os quais representam repetições no discurso, estes extraídos por código próprio[5]. Além dos grafos de fala, foram calculados índices léxico-semânticos que representam nossos corpora entre as línguas, estas, extraídas usando LexicalRichness[6], TextDescriptives[7] e Spacy[8]; alguns índices lexicais foram calculados manualmente.

[1] O conjunto de dados foi aprovado pelo Comitê de Ética da Universidade de Brasília (UnB)
[2] https://ffmpeg.org/
[3] https://github.com/openai/whisper
[4] https://github.com/pyannote/pyannote-audio
[5] Baseado em https://github.com/guillermodoghel/speechgraph
[6] https://lexicalrichness.readthedocs.io/en/
[7] https://github.com/HLasse/TextDescriptives
[8] https://spacy.io/

Adicionalmente, representações acústicas extraídas tanto por técnicas manuais quanto por técnicas baseadas em *embeddings* provenientes de redes neurais foram utilizadas. Para extrair representações manuais, utilizaram-se os conjuntos ComParE e eGeMAPS extraídos com a biblioteca OpenSmile[9] e o conjunto pAA[10]. Foram utilizados Wav2Vec2, VGGish[11], TRILL[12] como *embeddings*. Para o conjunto Wav2Vec2, utilizaram-se os modelos com ajuste fino para inglês[13] e português[14]. Após a extração, foi calculada a média dos vetores de saída dos modelos.

A disfluência foi representada com um algoritmo próprio que concatena três abordagens. A primeira utiliza o modelo de diarização Pyannote para calcular o tempo entre as sentenças e, assim, medir a contagem e a duração total das pausas. A segunda utiliza a biblioteca Calpy[15] e a terceira é calculada com a biblioteca Librosa[16].A dimensionalidade das representações geradas pode ser visualizada na Tabela 1.

Tabela 1. Dimensionalidade dos Conjuntos de Características

Conjunto	eGeMAPS	ComParE	pAA	wav2vec	trill	VGGish	vad	Lexical	SpeechGraph
Dimen.	88	6373	136	1028	2048	128	6	57	12

Para investigar o comportamento de cada conjunto de características agnósticas aos idiomas, avaliou-se os modelos treinados com os dados em inglês no conjunto de dados em português. Adotou-se a fusão tardia combinando os resultados de modelos treinados independentemente pelo método de votação suave. Essa técnica trata a probabilidade de classe de cada modelo como um voto ponderado, e a predição final é derivada da média das probabilidades de todos os modelos. Ao utilizar a fusão tardia com votação suave, pôde-se avaliar as sinergias entre os conjuntos de características, testando exaustivamente todas as 2.555 combinações possíveis.

Para os experimentos, foram selecionados os algoritmos de classificação Árvore de Decisão (DT), Random Forest, XGBoost, LightGBM e CatBoost. Foi aplicado *gridsearch* no conjunto de treino para ajustar os hiperparâmetros. Contudo, essa etapa apresentou limitações, pois hiperparâmetros otimizados para a língua de origem ocasionalmente degradavam o desempenho na língua de destino. Para cada conjunto de características, comparou-se o F1-score do modelo otimizado pelo *gridsearch* com o modelo com hiperparâmetros padrão; quando o desempenho padrão era superior na língua de teste, ele foi adotado como configuração final. Os melhores resultados de cada modelo estão apresentados na Tabela.2.

O algoritmo LightGBM alcançou 0,8000 de acurácia e 0,7995 de F1-score junto com a DT que precisou de menos conjuntos (Índices Lexicais, Trill e ComParE). Índices Lexicais e o ComParE estão presentes nos três melhores modelos. Todos os algoritmos apresentaram desempenhos similares, com a DT alcançando uma acurácia competitiva de 0,8000. A falta de um conjunto de validação para realizar o *gridsearch* nos idiomas de teste pode influenciar os resultados; com um conjunto de validação, seria esperado que os algoritmos de *boosting* tivessem um desempenho superior. Outra possível explicação

[9] https://www.audeering.com/research/opensmile/

[10] https://github.com/tyiannak/pyAudioAnalysis

[11] https://www.kaggle.com/models/google/vggish/TensorFlow2/vggish/1

[12] https://www.kaggle.com/models/google/nonsemantic-speech-benchmark/tensorFlow2/trill-distilled/3

[13] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english

[14] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-portuguese

[15] https://github.com/YvonneYYu/calpy

[16] https://librosa.org/

Tabela 2. Dois Melhores Desempenhos por Modelo.

Modelo	Conjunto de Características	Acurácia	F1
XGBoost	vad, Lexical, vggish, wav2vec2	0,8000	0,7980
XGBoost	vad, Lexical, pAA, vggish, ComParE	0,8000	0,7980
LightGBM	**Lexical, pAA, eGeMAPSv02, ComParE**	**0,8000**	**0,7995**
LightGBM	**vad, Lexical, pAA, vggish, trill, ComParE**	**0,8000**	**0,7995**
CatBoost	vad, vggish, trill, ComParE	0,7750	0,7737
CatBoost	pAA, vggish, trill, ComParE	0,7750	0,7714
RF	pAA, vggish, wav2vec2	0,7500	0,7500
RF	vggish	0,7500	0,7494
DT	**Lexical, trill, ComParE**	**0,8000**	**0,7995**
DT	vggish, trill, ComParE	0,7750	0,7714

é que algoritmos de *boosting* requerem conjuntos de dados maiores para generalizar bem. Com apenas 225 amostras para treinamento, a DT pode ser mais eficaz em identificar os padrões.

As melhores combinações de características variam de acordo com o modelo; contudo, das características textuais, os Índices Lexicais se mantiveram presentes nas melhores combinações, indicando uma boa representação entre o inglês e o português. Entre representações acústicas, o conjunto ComParE demonstrou presença entre as melhores combinações, apesar de sua grande dimensionalidade. Nota-se que o modelo LightGBM treinado apenas com os índices textuais e extrações manuais de características acústicas (*Lexical, pAA, eGeMAPSv02, ComParE*) teve desempenho igual ou superior a outros modelos que utilizam *embeddings*, indicando que uma arquitetura simplificada pode atingir os mesmos valores que uma mais complexa.

3. Conclusão

Este trabalho demonstrou a viabilidade de uma abordagem de classificação agnóstica ao idioma para a detecção da Doença de Alzheimer. A arquitetura proposta vai desde a extração de características até o treinamento dos modelos, utilizando características acústicas e textuais comuns ao inglês e ao português brasileiro.

A arquitetura mostrou-se promissora em distinguir indivíduos com DA de controles saudáveis. Contudo, o baixo número de amostras nos dados em português (40) leva a métricas sensíveis em que o acerto ou erro de poucas predições altera os resultados de forma significativa. Os valores de acurácia idênticos (ex.: 0,8000) entre os diferentes modelos podem ser um indicativo desse problema. Para mitigar esse problema, em trabalhos futuros pretende-se explorar a aplicação de técnicas de aumento de dados, assim como o uso de algoritmos de aprendizado profundo na classificação. Se dados em outros idiomas se tornarem disponíveis, pode-se expandir a análise para incluí-los.

Referências

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio.

Luz, S., Haider, F., Fromm, D., Lazarou, I., Kompatsiaris, I., and MacWhinney, B. (2024). An overview of the adress-m signal processing grand challenge on multilingual alzheimer's dementia recognition through spontaneous speech. *IEEE Open Journal of Signal Processing*, 5:738–749.

Vigo, I., Coelho, L., and Reis, S. (2022). Speech- and language-based classification of alzheimer's disease: A systematic review. *Bioengineering*, 9(1):27.