Segmentação Automática de Áudios na Entrevista de Pacientes para o Teste de Alzheimer

Gustavo Amaral[1], Luan Dopke[1], João Paulo Aires[1], Juliana Onofre de Lira [2], Lilian Cristine Hubner[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

{g.losch, luan.dopke}@edu.pucrs.br, dalvan.griebler@pucrs.br

Resumo. *A análise automática de áudios de pacientes tem-se mostrado promissora na detecção de condições neurológicas. Entretanto, a segmentação manual das falas entre entrevistador e entrevistado é um processo demorado e sujeito a erros. Este artigo propõe um pipeline automatizado para segmentar trechos de fala entre entrevistadores e entrevistados em entrevistas do teste Boston Diagnostic Aphasia Examination (BDAE) em português, combinando representações vetoriais e modelos de Aprendizado de Máquina supervisionados e não supervisionados. Nos experimentos, o modelo supervisionado XGBoost alcançou 0,8861 de acurácia, enquanto o modelo não supervisionado K-Means obteve 0,8458.*

1. Introdução

Entrevistas gravadas, como no teste *Cookie Theft* do *Boston Diagnostic Aphasia Examination* [Goodglass et al. 2001] (BDAE), utilizado no diagnóstico de doenças neurocognitivas, podem fornecer informações relevantes para análise linguística e acústica voltadas à detecção da Doença de Alzheimer (DA). A aplicação de técnicas de Aprendizado de Máquina tem se mostrado promissora nesse contexto [Vigo et al. 2022], especialmente para a detecção precoce, que é essencial para um tratamento eficaz e para retardar a progressão da doença. Soluções automatizadas, portanto, podem simplificar e agilizar a extração de características que auxiliam o profissional da saúde na obtenção de diagnósticos mais precisos. Na literatura, observam-se abordagens que exploram técnicas de segmentação em diferentes domínios. Por exemplo, no campo educacional, [García et al. 2024] aplica aprendizado de máquina supervisionado para identificar trechos de práticas de ensino momento da aula. No entanto, a dependência de anotações manuais limita a aplicabilidade desses métodos em cenários clínicos, onde a anotação manual de dados é cara e demorada.

Diante disso, este trabalho propõe e avalia diferentes técnicas para a segmentação dos áudios nas classes entrevistador e entrevistado em entrevistas do BDAE utilizando aprendizado de máquina. São abordados tanto modelos de aprendizado supervisionado quanto modelos não supervisionados, explorando também diferentes estratégias de pré-processamento e representação dos dados, como ilustrado na Figura 2.

2. Metodologia e Desenvolvimento

O conjunto de dados utilizado neste estudo contém 161 gravações de entrevistas em português brasileiro do teste *Cookie Theft*, divididas nas seguintes classes: sem diagnóstico de demência, DA, comprometimento cognitivo leve, demência semântica, demência vascular, diagnóstico não especificado no momento e demência mista. Em

cada gravação, um entrevistador solicita ao entrevistado que descreva verbalmente com o máximo detalhamento a imagem *Cookie Theft* (Figura 1). Entretanto, para que esse material seja utilizado em modelos preditivos, é necessário segmentar os trechos de áudio falados pelo entrevistador e pelo entrevistado. Dessa forma, é possível treinar modelos utilizando somente os intervalos de áudio falados pelo entrevistado, evitando o processamento de segmentos de áudio não relevantes para a análise.

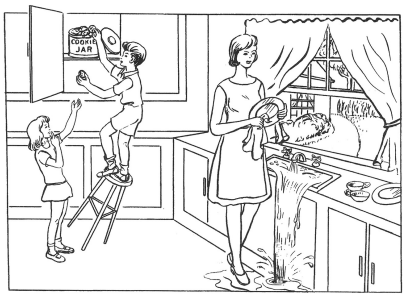

Figura 1. Cookie Theft.

Além dos métodos supervisionados e não supervisionados, foram exploradas duas abordagens para representar os dados, sendo a primeira a representação vetorial das transcrições textuais das entrevistas. O *pipeline* desenvolvido inicia com a transcrição e diarização dos áudios, etapa na qual se obtém o conteúdo falado em formato textual, juntamente com os instantes de tempo de cada palavra e a atribuição de segmentos falados a diferentes locutores. Para isso, utilizou-se o modelo WhisperX[1], que integra o Whisper, responsável pelo Reconhecimento Automático de Fala (ASR) e o modelo PyAnnote[2], estado da arte utilizado na diarização do áudio.

Com as transcrições geradas, foram aplicados procedimentos de pré-processamento com o objetivo de avaliar se a remoção de frases curtas impactaria no desempenho da segmentação. Análises exploratórias indicaram que uma quantidade significativa de frases possuía três ou menos palavras, geralmente compostas por cumprimentos ou interjeições, que pouco contribuíam para a tarefa proposta. Com isso, três configurações foram testadas: remoção de frases com até três palavras, remoção com até duas palavras e manutenção de todas as frases. Entre essas, a remoção de frases com até duas palavras apresentou o melhor equilíbrio entre desempenho e preservação do volume de dados.

A primeira abordagem consiste na geração de representações vetoriais textuais das transcrições geradas. Para representar vetorialmente o conteúdo textual, empregou-se o modelo *jina-embeddings-v3*[3], que apresentou desempenho expressivo no *MTEB*[4] (*Massive Text Embedding Benchmark*) com 58,37 pontos *MTEB*. Esse modelo transforma cada trecho transcrito em um vetor de 1024 dimensões, capturando relações semânticas presentes na linguagem natural. Com o objetivo de otimizar o processamento e reduzir a dimensionalidade dessas representações, aplicou-se o algoritmo de Análise de Componentes Principais (PCA), reduzindo a dimensionalidade dos vetores de 1024 para 380 componentes. A técnica obteve o melhor resultado com o modelo XGBoost, atingindo 0,8861 de acurácia, conforme mostrado na Tabela 1.

Também foi avaliada a representação vetorial de áudio, com o objetivo de capturar características acústicas da fala dos locutores, como o timbre e a entonação da voz. Para isso, após a etapa de transcrição, foram geradas representações de 1024 dimensões para cada enunciado usando o modelo *Wav2Vec 2*[5] com ajuste fino para a língua portuguesa. Foi utilizado o PCA para reduzir a dimensionalidade dos vetores para 380 componentes. Contudo, os resultados dessa abordagem foram inferiores aos obtidos com a representação vetorial textual, chegando ao máximo em 0,5419 de acurácia. Em ambas as técnicas de

[1]*WhisperX*: `https://github.com/m-bain/whisperX`
[2]*PyAnnote*: `https://huggingface.co/pyannote/speaker-diarization-3.1`
[3]*jina-embeddings-v3*: `https://huggingface.co/jinaai/jina-embeddings-v3`
[4]Link para acesso ao *MTEB*: `https://huggingface.co/spaces/mteb/leaderboard`
[5]*Wav2Vec 2*: `https://huggingface.co/facebook/wav2vec2-large-xlsr-53-portuguese`

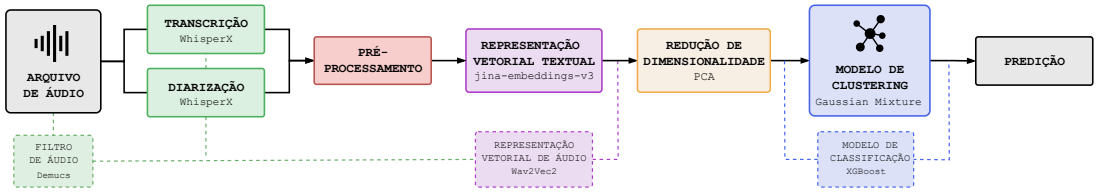

Figura 2. Diagrama do fluxo de dados, componentes tracejados foram testados mas não utilizados na arquitetura final.

representação vetorial, avaliou-se a remoção de ruído com o modelo *Demucs*[6], mas não houve ganho de desempenho na transcrição, com o WER (Word Error Rate) inalterado, e a acurácia da segmentação diminuiu, levando à exclusão dessa etapa do *pipeline* final.

Na etapa de segmentação dos locutores por aprendizado de máquina, foram explorados os métodos supervisionados e não supervisionados. Apesar da presença de anotações manuais, sua quantidade é limitada e o custo de rotulagem é elevado, justificando a inclusão de métodos não supervisionados. Dessa forma, optou-se por incluir experimentos em arquiteturas que não dependem de anotações prévias, possibilitando que o fluxo proposto seja implementado em dados não anotados no futuro. Para o treinamento e avaliação dos modelos, foram utilizadas as representações vetoriais textuais e acústicas, ambas com a dimensionalidade reduzida para 380 dimensões pelo algoritmo PCA.

Nos métodos não supervisionados, testou-se os algoritmos *Gaussian Mixture*, *K-Means* e *Agglomerative Clustering* limitando a saída dos modelos às classes entrevistador e entrevistado. Para os métodos supervisionados, foram avaliados os algoritmos *Random Forest* e *XGBoost*, treinados com o conjunto de dados usando as anotações, como os intervalos de tempo de fala do entrevistado, do entrevistador e suas transcrições.

Como comparação, propôs-se uma abordagem alternativa denominada *Locutor Predominante*, na qual o entrevistado é assumido como o locutor com maior tempo de fala, identificado a partir dos instantes de início e término de cada locutor na transcrição. Entretanto, essa abordagem apresenta resultados inferiores à segmentação por aprendizado de máquina, atingindo 0,6214 de acurácia, conforme mostrado na Tabela 1.

A avaliação das arquiteturas testadas foi feita com base na comparação entre as previsões geradas pelos modelos e as anotações manuais, calculando métricas como acurácia, F1-Score e o número de linhas após a etapa de pré-processamento. As métricas para o aprendizado supervisionado foram calculadas em um conjunto de teste representado por 33% dos dados do conjunto total de dados. Observou-se que a remoção de frases curtas aumenta a acurácia, mas também reduz a quantidade de informação disponível. Esse *trade-off* evidencia a importância de equilibrar a etapa de pré-processamento dos dados com a preservação de informações relevantes, especialmente porque indivíduos com DA tendem a falar menos, e a exclusão de suas falas pode introduzir viés nos resultados.

3. Conclusões

Os experimentos mostraram que a combinação de transcrição, diarização e representações semânticas textuais permite segmentar entrevistadores e entrevistados com bom desempenho. O modelo supervisionado *XGBoost*, com o pré-processamento retirando frases com até 3 palavras, obteve o melhor desempenho, atingindo 0,8861 de acurácia, mas exige anotação de dados, restringindo sua aplicação a contextos de dados anotados. Já o modelo não supervisionado *K-Means*, com o pré-processamento retirando frases com até 3 palavras, alcançou desempenho competitivo, atingindo 0,8458 de acurácia em seu melhor resultado, tornando-se adequado para cenários onde não há a possibilidade de anotação dos dados.

[6]*Demucs*: `https://github.com/adefossez/demucs`

Tabela 1. Tabela de resultados em diferentes combinações de arquiteturas agrupados por estratégia de pré-processamento. Em negrito as melhores combinações por agrupamento e método de aprendizado. N.a.: Não aplicável.

Modelos	Pré-Processamento	PCA	Acurácia	F1-Score	Total de Linhas
K-Means	Não	380	0,6985	0,7179	2834
Agglomerative Cluster	Não	380	0,5860	0,5486	2834
Gaussian Mixture	**Não**	**380**	**0,7700**	**0,7677**	**2834**
Random Forest	Não	N.a.	0,8386	0,8383	2834
XGBoost	**Não**	**N.a.**	**0,8621**	**0,8621**	**2834**
K-Means	Sim (remove $\leq$2 palavras)	380	0,7709	0,7660	2131
Agglomerative Cluster	Sim (remove $\leq$2 palavras)	380	0,7283	0,7565	2131
Gaussian Mixture	**Sim (remove $\leq$2 palavras)**	**380**	**0,7903**	**0,7972**	**2131**
Random Forest	Sim (remove $\leq$2 palavras)	N.a.	0,8267	0,8235	2131
XGBoost	**Sim (remove $\leq$2 palavras)**	**N.a.**	**0,8650**	**0,8649**	**2131**
K-Means	**Sim (remove $\leq$3 palavras)**	**380**	**0,8458**	**0,8439**	**1861**
Agglomerative Cluster	Sim (remove $\leq$3 palavras)	380	0,7748	0,7619	1861
Gaussian Mixture	Sim (remove $\leq$3 palavras)	380	0,8001	0,7992	1861
Random Forest	Sim (remove $\leq$3 palavras)	N.a.	0,8520	0,8476	1861
XGBoost	**Sim (remove $\leq$3 palavras)**	**N.a.**	**0,8861**	**0,8859**	**1861**
Locutor Predominante	N.a.	N.a.	0,6214	0,5456	2834

O pré-processamento, especialmente a remoção de frases curtas, mostrou impacto significativo na melhoria da acurácia, embora possa retirar informações possivelmente importantes presentes nas frases curtas do conjunto de dados. Atualmente, o trabalho apresenta algumas limitações que podem ser abordadas em pesquisas futuras, como a aplicação de algoritmos de otimização de parâmetros nos modelos e métodos testados, bem como a implementação de uma etapa de pré-processamento para remover frases curtas na representação vetorial de áudio. Experimentos futuros também poderão investigar a avaliação de diferentes modelos de representação vetorial, tanto textual quanto acústica,, além de explorar de forma mais detalhada as propriedades acústicas, incluindo a extração e análise de *features* manuais, como o conjunto ComParE [Weninger et al. 2013].

Portanto, a abordagem proposta baseada no modelo *Gaussian Mixture* com representação vetorial textual e pré-processamento removendo frases com até 2 palavras, se enquadra como a mais adequada, pois combina um bom desempenho sem a necessidade de depender de dados anotados. Esse fluxo pode ainda ser implementado como etapa de pré-processamento ou limpeza em estudos futuros que utilizem conjuntos de dados de Alzheimer sem anotação, proporcionando uma base robusta para o treinamento de modelos de aprendizado de máquina voltados à detecção da doença.

Referências

García, F. P., Cánovas, O., and Clemente, F. J. G. (2024). Exploring ai techniques for generalizable teaching practice identification. *IEEE Access*, 12:134702–134713.

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *Boston Diagnostic Aphasia Examination: Stimulus Cards*. Lippincott Williams & Wilkins.

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

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., and Scherer, K. R. (2013). On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in Psychology*, Volume 4 - 2013.