

Enriquecimento de Embeddings Multimodais com Representações de Encoders de Fala para Pontuação Automática de Fluência de Leitura

Hyan H. N. Batista¹, André C. A. Nascimento¹,
Péricles Miranda¹, Rafael Ferreira Mello¹

¹Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brasil

{hyan.batista, andre.camara, pericles.miranda, rafael.mello}@ufrpe.br

Resumo. A avaliação automática da fluência de leitura é um problema relevante para o apoio a contextos educacionais, mas ainda há uma lacuna na literatura quanto a comparações sistemáticas entre representações pré-treinadas, de uma discussão clara sobre a formulação adequada da tarefa e de análises críticas sobre vieses demográficos. Neste trabalho, avaliamos quatro grupos de embeddings multimodais (Wav2Vec2, Whisper, BERT e Gemini Embedding 2, isolados ou combinados) e oito regressores ensemble no dataset SpeechOcean762 para realizar um benchmark de fluência. Os resultados mostram que a configuração Whisper + Gemini + PCA-256 com Stacking/Voting alcança melhores resultados com PCC de 0,77 e Kappa de Cohen de 0,76. Além disso, foi realizado um estudo de viés demográfico por gênero e faixa etária para avaliar possíveis vieses nos modelos.

Abstract. Automatic assessment of reading fluency is a relevant problem for educational contexts, but there remains a gap in the literature regarding systematic comparisons of pre-trained representations, a clear discussion of appropriate task formulation, and critical analyses of demographic biases. In this work, we evaluate four groups of multimodal embeddings (Wav2Vec2, Whisper, BERT, and Gemini Embedding 2, either in isolation or combined) and eight ensemble regressors on SpeechOcean762 to establish a benchmark for fluency. The results show that the Whisper + Gemini + PCA-256 configuration with Stacking/Voting achieves the best results, with a PCC of 0.77 and Cohen's Kappa of 0.76. Additionally, a demographic bias study by gender and age group was conducted to assess potential biases in the models.

1. Introdução

A fluência de leitura é um indicador central do desenvolvimento de habilidades linguísticas e está fortemente associada ao desempenho escolar e à compreensão textual [van der Velde et al. 2025, van der Velde et al. 2024]. Contudo, a avaliação exige tempo do educador e apresenta grande variabilidade quando diferentes professores a realizam [Bai et al. 2021]. Esses fatores dificultam a avaliação em larga escala, especialmente em contextos educacionais com alta razão aluno/professor, o que motiva a busca por ferramentas automáticas capazes de produzir escores consistentes a partir do áudio da leitura [da Silva et al. 2025].

O avanço recente de modelos pré-treinados de fala (como Wav2Vec2 e Whisper) e de linguagem (como BERT) tornou viável extrair, de forma direta, representações vetoriais (*embeddings*) que capturam propriedades acústico-fonéticas e semânticas relevantes para a tarefa [da Silva et al. 2025, Liu et al. 2023b, Mohammadi et al. 2025]. Mais recentemente, modelos multimodais de larga escala, a exemplo do Gemini Embedding 2, passaram a oferecer *embeddings* que integram áudio e texto em um único espaço vetorial, abrindo a possibilidade de combinar sinais complementares sem a necessidade de arquiteturas ad hoc de fusão [Liu et al. 2023a]. Apesar desse avanço, a literatura sobre pontuação automática de fluência de leitura ainda carece de comparações sistemáticas entre diferentes *embeddings*, de uma discussão clara sobre a formulação adequada da tarefa (classificação em níveis discretos versus regressão contínua) e de análises críticas sobre o viés demográfico dos preditores resultantes, uma dimensão particularmente sensível em aplicações educacionais.

Neste trabalho, investigamos empiricamente, utilizando o conjunto público *SpeechOcean762*, a capacidade preditiva de quatro grupos de *embeddings* (Wav2Vec2, Whisper, BERT e Gemini Embedding 2), isolados ou combinados, para as quatro notas anotadas por avaliadores humanos: acurácia, completude, fluência e prosódia. Os resultados mostram que a configuração baseada na fusão de Whisper e Gemini, com redução de dimensionalidade por PCA-256 e regressão por *ensembles* (*Stacking* e *Voting*), atinge correlação de Pearson de até 0,77 e Kappa de Cohen de até 0,76 em relação a avaliação humana.

As contribuições deste trabalho são: (i) uma comparação sistemática entre quatro grupos de *embeddings* multimodais e oito regressores *ensemble* para a pontuação automática de fluência de leitura; (ii) um método proposto baseado em *Whisper + Gemini + PCA-256*, que estabelece um novo patamar de desempenho sobre o *SpeechOcean762* entre as combinações avaliadas; e (iii) uma análise de viés demográfico, estratificada por gênero e faixa etária, que identifica padrões sistemáticos de erro associados, sobretudo, à distribuição dos dados de treino, com implicações diretas para o uso responsável desses modelos em contextos educacionais.

2. Embasamento teórico

2.1. Fluência de Leitura

O conceito de Fluência de Leitura Oral (em inglês, *Oral Reading Fluency* - ORF) combina três elementos [da Silva et al. 2025]: acurácia, velocidade e prosódia. A ORF pode ser definida, portanto, como a habilidade de ler precisamente a uma velocidade apropriada, com prosódia adequada [Hasbrouck and Glaser 2012].

O primeiro deles é a acurácia. Esse elemento trata da capacidade do leitor de decodificar e pronunciar as palavras corretamente, sem cometer erros [Wang et al. 2025]. Ele depende do conhecimento das letras e de habilidades fonológicas, como a consciência fonêmica. Segundo [Hayati and Saryanto 2025], a consciência fonêmica é a habilidade de reconhecer e manipular fonemas, sendo um forte preditor do desenvolvimento de habilidades de decodificação e do sucesso na aprendizagem. Os autores definem decodificação como o entendimento das correspondências letra-som e o uso desse conhecimento para ler palavras conhecidas e desconhecidas. Além disso, dificuldades de fonética e de decodificação podem estar associadas a problemas de ORF.

Tão importante quanto a acurácia é o segundo elemento que compõe o conceito de ORF: a velocidade ou automaticidade. A Teoria da Automaticidade, amplamente desenvolvida por [LaBerge and Samuels 1974], propõe que a fluência de leitura é obtida por meio do reconhecimento automático de palavras, permitindo que o leitor concentre sua habilidade cognitiva na compreensão. A importância dessa métrica é tal que alguns trabalhos, como o de [van der Velde et al. 2025] chegam a indicar que a velocidade é frequentemente o principal fator para distinguir leitores fluentes de não fluentes nos anos iniciais do ensino fundamental.

Por fim, o terceiro e último elemento, a prosódia, pode ser entendido como o elemento expressivo da leitura, que inclui o uso adequado de ritmo, entonação, fraseamento, pausas e estresse vocal [Harmsen et al. 2025]. Uma boa prosódia é um forte indicativo de um bom leitor, pois, como coloca [Bailly et al. 2022], um leitor especialista é descrito como alguém que lê como fala, que não está apenas decodificando com precisão, em uma taxa conversacional, mas também usando melodia, ritmo e variações de intensidade para capturar a atenção dos ouvintes e facilitar a compreensão do texto. Ainda segundo os autores, portanto, um bom leitor refere-se não apenas à acurácia e à automaticidade, mas, em sua maior parte, ao uso de sua habilidade para aplicar características prosódicas para renderizar o significado do texto lido em voz alta.

2.2. Trabalhos Relacionados

Embora a automação do processo de avaliação da leitura oral acelere o fornecimento e o recebimento de feedback, diagnósticos detalhados oferecem benefícios adicionais ao processo de ensino e aprendizagem. Nesse sentido, estudos recentes têm investigado as capacidades de *Automatic Speech Recognition* (ASR) para avaliar a ORF como um construto multidimensional composto por acurácia, velocidade (automaticidade) e prosódia [Wang et al. 2025, da Silva et al. 2025, van der Velde et al. 2024]. No que diz respeito à avaliação global, tradicionalmente medida por meio de *Words Correct Per Minute* (WCPM), pesquisas iniciais demonstraram que avaliações automatizadas poderiam se aproximar fortemente da avaliação humana em termos de confiabilidade [da Silva et al. 2025, Harmsen et al. 2025, Yıldız et al. 2025].

A partir de uma perspectiva mais complexa, pesquisas também têm explorado a efetividade de *pretrained foundation models* na captura de comportamentos de leitura mais sutis. Por exemplo, o Whisper (OpenAI) alcançou desempenho de estado da arte em nível de palavra, embora seu decodificador de linguagem tenda a “autocorrigir” erros de leitura, o que pode mascarar dificuldades genuínas [da Silva et al. 2025] [Gao et al. 2024]. Em contraste, modelos como Wav2Vec 2.0 e HuBERT utilizam *self-supervised learning* (SSL) para extrair características acústicas em nível de frame, mostrando-se superiores em *recall* na detecção de *miscues* específicos de leitura, como substituições e inserções, pois preservam variações fonéticas [Gao et al. 2024]. Particularmente no contexto de aprendizes não nativos, estudos com o dataset Speechocean762 utilizaram características SSL para descobrir unidades acústicas sem a necessidade de um texto de referência, alcançando desempenho comparável ao humano em cenários de “*read aloud*” [Liu et al. 2023b].

Além disso, diversos estudos investigaram abordagens multimodais para melhorar a estabilidade das avaliações de fluência. [Liu et al. 2023a] propuseram um método que

combina a última *hidden layer* de modelos Wav2Vec 2.0 *fine-tuned* com características textuais decodificadas por meio de mecanismos de *cross-modal attention*, resultando em avaliações mais confiáveis do que as de modelos baseados apenas em áudio. De forma semelhante, arquiteturas de aprendizado profundo, como ResNet e BERT, têm sido exploradas para a classificação de proficiência em L2, com modelos BERT melhorando significativamente a acurácia da classificação baseada em texto, mesmo em contextos com recursos limitados [Mohammadi et al. 2025].

Recentemente, há um grande foco em modelos de ensemble, como Random Forest, XGBoost e Logistic Regression, para mapear características extraídas no contexto de fluência de leitura [da Silva et al. 2025]. Achados recentes indicam que modelos de regressão superam consistentemente a classificação discreta na explicação da variabilidade das avaliações especializadas de fluência, confirmando que a fluência em leitura é uma tarefa inerentemente contínua e ordinal [da Silva et al. 2025].

No entanto, várias limitações persistem na literatura atual. Primeiramente, sobre uso de modelos multimodais de larga escala. O surgimento recente do *multimodal embedding* desenvolvido por [?], como a Gemini Embedding 2, por exemplo, evidencia a necessidade de avaliar sua eficácia específica na avaliação multimodal de fluência. Uma avaliação abrangente que integre essas representações multimodais avançadas a técnicas de regressão regularizada é crucial para alcançar diagnósticos educacionais mais equitativos e detalhados.

Outra limitação da literatura diz respeito às disparidades raciais e de dialetos/sotaques com evidências sugerindo que avaliadores humanos e sistemas ASR podem apresentar diferentes vieses sistemáticos ao avaliar estudantes de grupos marginalizados [Petscher et al. 2026]. Para além disso, os vieses de gênero e de idade em modelos de ASR focados na fluência de leitura permanecem subexplorados.

2.3. Questões de Pesquisa

A partir das lacunas observadas na literatura e dos artefatos experimentais produzidos neste estudo, formulamos duas questões de pesquisa que orientam a investigação.

QUESTÃO DE PESQUISA 1 (QP1):

Qual a capacidade de representações (embeddings) multimodais pré-treinadas em modelos de fala e linguagem (Wav2Vec2, Whisper, BERT e Gemini) de prever automaticamente os escores de fluência de leitura atribuídos por avaliadores humanos?

QUESTÃO DE PESQUISA 2 (QP2):

Qual o desempenho dos modelos de regressão em diferentes grupos demográficos (gênero e faixa etária) e até que ponto os preditores apresentam vies sistemático entre esses grupos?

3. Método

3.1. Dados e Pré-processamento

Utilizamos o conjunto de dados *SpeechOcean762*, um *corpus* público distribuído via *HuggingFace* que contém 5.000 enunciados de leitura em inglês de aprendizes não nativos,

divididos de forma fixa em 2.500 amostras de treinamento e 2.500 de teste. Cada amostra contém o sinal de áudio bruto, a transcrição esperada da passagem, identificadores do locutor e atributos demográficos (gênero e idade), além de quatro escores atribuídos por avaliadores humanos na escala de 0 a 10: acurácia, completude, fluência e prosódia.

Adotamos a divisão original de treino e teste sem reamostragem ou validação cruzada, a fim de manter a comparabilidade com trabalhos que empregam o mesmo *corpus*. Os áudios são processados em sua forma bruta, e a normalização e a tokenização são realizadas pelos próprios processadores dos modelos pré-treinados. As transcrições textuais são fornecidas aos modelos de linguagem como entrada bruta, sem normalização adicional.

3.2. Embeddings

Foram considerados quatro modelos pré-treinados para extração de representações vetoriais a partir das amostras de áudio e de seus transcritos:

- **Wav2Vec2** [Baevski et al. 2020] (`facebook/wav2vec2-base`): produz representações acústicas de 768 dimensões obtidas pela média temporal das ativações do *encoder* convolucional-transformer.
- **Whisper** [Radford et al. 2023] (`openai/whisper-base`): a saída do *encoder* é agregada produzindo um vetor de 1.280 dimensões, capturando informação acústico-fonética e prosódica.
- **BERT** [Devlin et al. 2019] (`bert-base-uncased`): aplicado sobre a transcrição da passagem lida, fornecendo um vetor textual de 768 dimensões.
- **Gemini Embedding 2** [?]: modelo multimodal acessado via API, com `output_dimensionality=768`, que recebe conjuntamente o áudio e o texto da passagem e devolve um único vetor fundido de 768 dimensões.

A partir desses modelos compomos quatro grupos de características empregados nos experimentos: (a) *Wav2Vec + Texto* (concatenação de Wav2Vec2 e BERT, totalizando 1.536 dimensões); (b) *Whisper + Texto* (concatenação de Whisper e BERT, 2.048 dimensões); (c) *Gemini* (apenas o vetor multimodal de 768 dimensões); e (d) ***Whisper + Gemini*** (concatenação do *encoder* de Whisper com o vetor multimodal do Gemini, 2.048 dimensões), grupo que corresponde à configuração adotada como método proposto. Para todas as configurações de regressão avaliamos ainda variantes em que o vetor resultante é projetado para 256 componentes via *Principal Component Analysis* (PCA), com o objetivo de mitigar a alta dimensionalidade, atenuar a colinearidade entre componentes redundantes de Whisper e Gemini e regularizar os *ensembles*. A combinação *Whisper + Gemini + PCA-256* é a configuração central proposta neste trabalho; as demais são apresentadas como linhas de base.

3.3. Modelos de Regressão

Para investigar o desempenho dos modelos de regressão para cada combinação de grupo de características, foram avaliados oito modelos *ensemble*: *Random Forest*, *Extra Trees*, *Gradient Boosting*, *Histogram-based Gradient Boosting*, *Bagging*, *AdaBoost*, *Voting* e *Stacking*. A escolha dos modelos foi baseada, em parte, em [da Silva et al. 2025], com foco em algoritmos de *ensembles* e *boosting*, pois geralmente apresentam desempenho superior [Fernández-Delgado et al. 2014]. Os regressores são treinados diretamente na

escala contínua da nota, tendo como alvo o processo. Os hiperparâmetros são mantidos nos valores padrão da biblioteca, priorizando a comparação justa entre representações e formulações de tarefa em detrimento de ajustes específicos por modelo.

3.4. Análise de resultados

Para avaliar a QP1, foi conduzida uma experimentação no conjunto de teste fixo do *SpeechOcean762*, com métricas tradicionais da literatura de regressão: (i) o *Erro Quadrático Médio* (MSE), (ii) sua raiz (RMSE), (iii) o *Coefficiente de Correlação de Pearson* (PCC) entre os escores preditos e os escores humanos.

3.5. Análise de viés demográfico

Para responder à QP2, realizamos uma análise estratificada das previsões dos regressores sobre os atributos demográficos disponíveis no *SpeechOcean762*: gênero (masculino, feminino) e faixa etária, agrupados em cinco intervalos (6–10, 11–13, 14–16, 17–20 e 21+ anos). Para cada subgrupo computamos quatro métricas: RMSE, PCC, erro médio sinalizado ($\hat{y} - y$, em que valores positivos indicam sobrestimação sistemática) e erro absoluto médio. Definimos ainda a *diferença de desempenho* entre dois grupos como $\Delta = m_{\text{feminino}} - m_{\text{masculino}}$, de modo que valores positivos de Δ RMSE indicam pior desempenho do grupo feminino e valores positivos de Δ erro médio indicam maior sobrestimação do grupo feminino. A análise é conduzida para todos os quatro grupos de características e todos os oito regressores, permitindo verificar se eventuais vieses são específicos de um modelo ou pervasivos entre as arquiteturas, indicando origem nos dados de treino.

4. Resultados

Nesta seção, apresentamos os resultados de regressão para cada combinação (grupo de características $\times$ alvo) avaliada. Os valores reportados nas Tabelas 1 e 2 são obtidos a partir do conjunto de teste do *SpeechOcean762* e correspondem às métricas descritas em 3.4.

De modo geral, observamos padrões consistentes ao longo das tabelas. Primeiro, o alvo de completude é fortemente desbalanceado no *corpus*: praticamente todas as amostras recebem um resultado bem melhor do que nas outras categorias, o que faz com que as métricas de MSE e RMSE sejam muito melhores, enquanto o PCC ainda não é tão alto, indicando que o desempenho aparente decorre da classe majoritária e não de discriminação efetiva. Segundo, para os demais alvos (acurácia, fluência e prosódia), as métricas situam-se na faixa de aproximadamente 0,90 a 1,60 e os valores de MSE refletem a dificuldade de prever o resultado nessas categorias.

A configuração *Whisper + Gemini + PCA-256* apresentou melhor desempenho na maioria dos alvos de regressão, sendo a melhor combinação na Tabela 2 em fluência, prosódia e completude, além de obter os melhores MSE, RMSE e PCC em acurácia. Para o alvo de fluência, o regressor *Stacking* alcançou $\text{PCC} = 0,7742$ e $\kappa = 0,7567$, enquanto o *Voting* obteve o menor erro, com $\text{MSE} = 0,8192$ e $\text{RMSE} = 0,9051$. Para prosódia, o *Stacking* resultou em $\text{PCC} = 0,7667$ e $\kappa = 0,7436$, com $\text{MSE} = 0,8613$ no *Voting*. Para acurácia, o *Stacking* atingiu $\text{PCC} = 0,6716$ e $\kappa = 0,6202$, enquanto o *Voting* registrou $\text{MSE} = 1,3134$; nesse alvo, porém, o maior κ da Tabela 2 (0,6316) foi alcançado pela

configuração *Whisper + Texto + PCA-256* com *Stacking*. Em comparação, a melhor linha de base sem PCA-256 (*Wav2Vec + Texto* com *HistGradientBoosting*) atinge $PCC \approx 0,67$ e $\kappa \approx 0,61$ apenas em fluência, ficando bastante atrás em prosódia ($\kappa = 0,5842$) e em completude, e modestamente atrás em acurácia ($\kappa = 0,5784$).

Com relação à aplicação de PCA-256 (Tabela 2), *Whisper + Gemini + PCA-256* é a melhor combinação em 15 das 16 comparações ($4 \text{ alvos} \times 4 \text{ métricas}$). Ela lidera em todas as métricas de fluência ($\kappa = 0,7567$), prosódia ($\kappa = 0,7436$) e completude ($PCC = 0,0780$, único valor não trivial), além de obter os melhores MSE, RMSE e PCC em acurácia. A única exceção é κ em acurácia, onde *Whisper + Texto + PCA-256* alcança 0,6316 contra 0,6202. Os ganhos sobre *Wav2Vec + Texto + PCA-256* e *Gemini + PCA-256* são particularmente expressivos em fluência e prosódia, da ordem de 20 pontos percentuais em κ , indicando que o embedding multimodal Gemini preserva, mesmo após redução para 256 componentes, a informação relevante que a concatenação tradicional de *embeddings* acústicos e textuais não captura.

Tabela 1. Melhor regressor por combinação de embedding e alvo

Características	Alvo	Modelo	MSE	RMSE	PCC
Wav2Vec + Texto	Acurácia	HistGradientBoosting	1.4178	1.1907	0.6362
	Fluência	HistGradientBoosting	1.1273	1.0618	0.6670
	Prosódia	Stacking	1.3147	1.1466	0.6340
	Completude	ExtraTrees	0.2505	0.5005	0.3301
Gemini	Acurácia	HistGradientBoosting	1.6026	1.2659	0.5697
	Fluência	GradientBoosting	1.2974	1.1391	0.6046
	Prosódia	HistGradientBoosting	1.3005	1.1404	0.6063
	Completude	ExtraTrees	0.2629	0.5127	0.0075
Whisper + Gemini	Acurácia	HistGradientBoosting	1.2961	1.1385	0.6773
	Fluência	Stacking	0.9032	0.9503	0.7729
	Prosódia	Stacking	0.9346	0.9667	0.7557
	Completude	HistGradientBoosting	0.2631	0.5130	0.0099
Whisper + Texto	Acurácia	AdaBoost	1.2471	1.1167	0.6950
	Fluência	Stacking	0.9332	0.9660	0.7721
	Prosódia	Stacking	0.9584	0.9790	0.7596
	Completude	Bagging	0.2588	0.5087	0.1168

Tabela 2. Melhor regressor por combinação de embedding e alvo com PCA-256

Características	Alvo	Modelo	MSE	RMSE	PCC
Wav2Vec + Texto	Acurácia	HistGradientBoosting	1.5262	1.2354	0.6006
	Fluência	HistGradientBoosting	1.3291	1.1529	0.5891
	Prosódia	HistGradientBoosting	1.3690	1.1700	0.5790
	Compleitude	ExtraTrees	0.2628	0.5126	0.0146
Gemini	Acurácia	HistGradientBoosting	1.5789	1.2566	0.5816
	Fluência	HistGradientBoosting	1.2681	1.1261	0.6144
	Prosódia	HistGradientBoosting	1.2812	1.1319	0.6142
	Compleitude	ExtraTrees	0.2627	0.5126	0.0149
Whisper + Gemini	Acurácia	Stacking	1.3481	1.1611	0.6716
	Fluência	Stacking	0.8677	0.9315	0.7742
	Prosódia	Stacking	0.8732	0.9345	0.7667
	Compleitude	ExtraTrees	0.2615	0.5113	0.0780
Whisper + Texto	Acurácia	Stacking	1.4948	1.2226	0.6614
	Fluência	Stacking	1.0575	1.0284	0.7241
	Prosódia	Stacking	1.0878	1.0430	0.7073
	Compleitude	ExtraTrees	0.2629	0.5127	0.0079

4.1. Análise de viés por gênero e faixa etária

A análise estratificada descrita em 3.5 foi conduzida com base nas predições do regressor *HistGradientBoosting*, o modelo mais consistente entre os grupos de características, e, posteriormente, foi replicada em todos os grupos de características e de regressores. Antes de interpretar os erros, convém examinar a distribuição dos rótulos humanos por gênero, apresentada na Figura 1. As médias dos escores reais são ligeiramente superiores no grupo masculino para acurácia ($\mu_M = 7,79$ vs. $\mu_F = 7,59$), fluência (7,87 vs. 7,80) e prosódia (7,54 vs. 7,46); a completude é praticamente saturada em ambos os grupos ($\mu_M = 9,94$ vs. $\mu_F = 10,00$, com desvio padrão nulo no feminino).

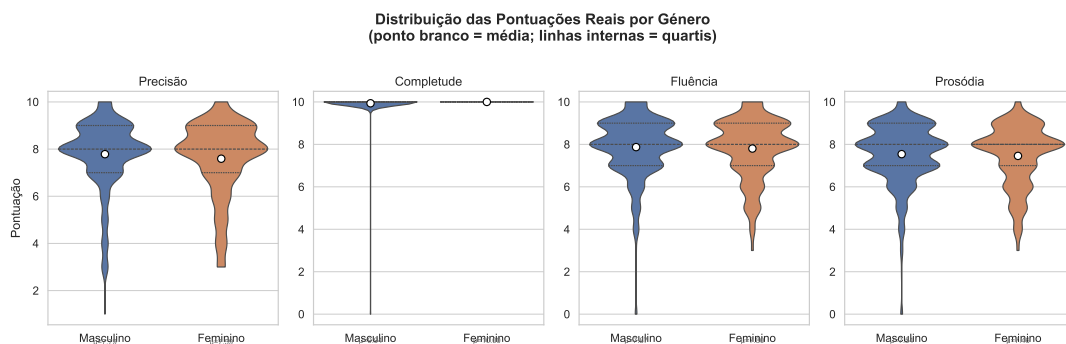


Figura 1. Distribuição dos escores reais do *SpeechOcean762* por gênero. O grupo masculino apresenta médias ligeiramente superiores em acurácia, fluência e prosódia.

A Figura 2 mostra a diferença de erro médio de predição entre os gêneros, $(\hat{y} - y)_F - (\hat{y} - y)_M$, para todas as combinações de grupos de características e regressores. Nos grupos baseados em BERT (*Wav2Vec + BERT* e *Whisper + BERT*), o modelo sobrestima sistematicamente o grupo masculino em acurácia, fluência e prosódia (valores negativos na figura), padrão que se inverte parcialmente nos grupos baseados em Gemini Embedding 2, conforme detalhado adiante. Em termos de RMSE, o modelo é consistentemente mais preciso para o grupo feminino em fluência e prosódia, com diferenças de

aproximadamente 0,08 a 0,18 em relação ao grupo masculino. Esse padrão de erro (o modelo ajusta-se melhor à faixa de pontuações ligeiramente mais baixas) é pervasivo entre os quatro grupos de características e a maioria dos regressores avaliados, o que sugere que o viés tem origem na distribuição dos dados de treino, e não em um algoritmo específico.

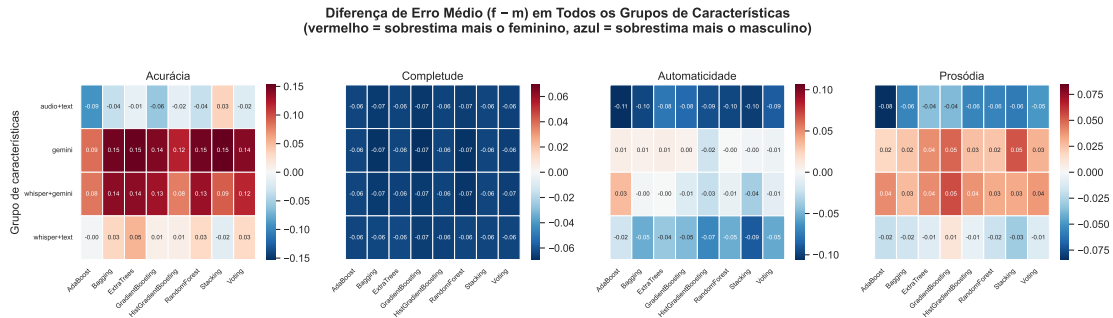


Figura 2. Diferença de erro médio de predição entre gêneros, $(\hat{y} - y)_F - (\hat{y} - y)_M$, para todos os grupos de características e regressores. Vermelho indica maior sobrestimação do grupo feminino; azul indica maior sobrestimação do grupo masculino. O padrão predominantemente azul nos grupos baseados em BERT confirma a sobrestimação sistemática do grupo masculino nessas configurações.

Os grupos de características baseados em *Gemini Embedding 2* puro e *Whisper + Gemini* introduzem, em contraste, um viés de acurácia favorável ao feminino (diferença de erro médio em torno de +0,13 para *Gemini* e +0,11 para *Whisper + Gemini*), com efeito de menor magnitude também em prosódia. Esse contraste em relação aos modelos locais (*Wav2Vec* e *Whisper* combinados com BERT) sugere que o modelo multimodal de larga escala captura aspectos linguísticos correlacionados com o gênero que não se manifestam nas representações acústico-textuais locais, como fica claro na Figura 2.

A Figura 3 cruza gênero e faixa etária para o regressor *HistGradientBoosting* no grupo *Wav2Vec + Texto*. O desempenho ótimo concentra-se na faixa de 14–16 anos (RMSE de 0,72 em fluência e 0,86 em prosódia) e degrada nas extremidades da distribuição etária: o modelo sobestima crianças de 6–10 anos (erro médio entre +0,11 e +0,18 nos três alvos não-saturados) e subestima a faixa de 14–20 anos (erro médio entre -0,13 e -0,35). A interação gênero $\times$ idade revela que o grupo masculino de 11–13 anos é o mais difícil de modelar (RMSE para fluência de 1,60), e a anomalia de completude concentra-se em masculinos jovens, com RMSE de 0,86 na faixa 6–10 e 1,43 na faixa 11–13 ao passo que, em todas as demais células, o RMSE de completude permanece próximo de 0,03.

5. Discussão

5.1. Capacidade de predição

Os resultados da Seção 4 indicam que *embeddings* pré-treinados de fala e linguagem capturam um sinal preditivo relevante para a fluência de leitura, mesmo sem ajuste fino para a tarefa. O método proposto, baseado em *Whisper + Gemini + PCA-256*, alcança correlações de Pearson de até 0,77 com os escores humanos para fluência e prosódia e

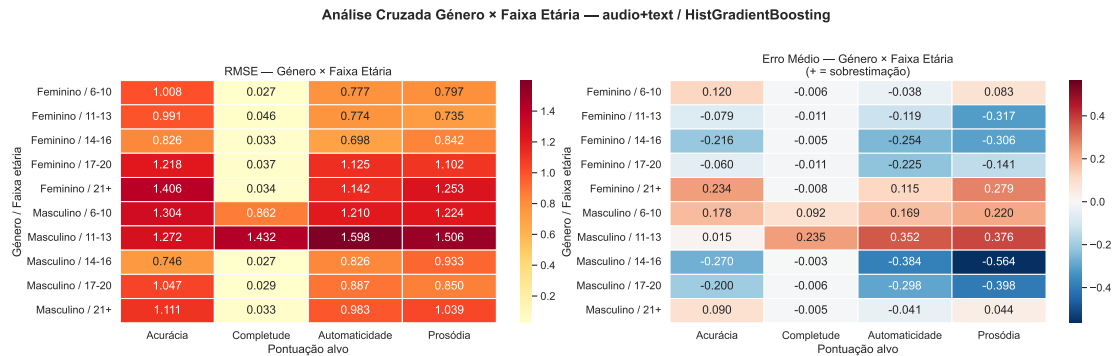


Figura 3. Análise cruzada gênero × faixa etária para o regressor *HistGradientBoosting* sobre *Wav2Vec* + *Texto*. Esquerda: RMSE; direita: erro médio de predição (positivo = sobrestimação).

de 0,67 para acurácia. A superioridade dessa configuração em relação às linhas de base baseadas em *Wav2Vec* + *BERT*, em *Whisper* + *BERT* ou em *Gemini* isolado evidencia que a complementaridade entre as representações acústico-fonéticas do *Whisper* e a fusão multimodal do *Gemini*, aliada à redução de dimensionalidade, é o fator decisivo de desempenho. Para mais, quando comparado aos resultados obtidos no trabalho original [Zhang et al. 2021], o método proposto apresenta uma melhora substancial tanto em PCC quanto MSE.

Entre os oito *ensembles* avaliados, *Stacking*, *Voting* e *HistGradientBoosting* foram consistentemente os mais competitivos em regressão, ao passo que modelos baseados em *bagging* de árvores (*Random Forest*, *Extra Trees*) ficaram um passo atrás nos alvos não-degenerados. A redução por PCA-256 teve, contudo, efeito heterogêneo: melhorou marginalmente as configurações baseadas em *Gemini* (*Gemini* puro e *Whisper* + *Gemini* em prosódia e completude), mas degradou de forma consistente as configurações que combinam *embeddings* acústicos com *BERT* (*Wav2Vec* + *BERT* e *Whisper* + *BERT*), com quedas de aproximadamente 3 a 9 pontos percentuais em κ . Esse padrão sugere que a redundância dimensional explorada pela projeção é particularmente acentuada nos *embeddings* multimodais densos do *Gemini*, em que a PCA atua simultaneamente como compressão e regularização, ao passo que a concatenação tradicional acústico-textual perde informação relevante durante a compressão.

5.2. Viés demográfico

A análise estratificada (Seção 4.1) revela três padrões relevantes. Primeiro, há *viés sistêmico* por gênero: o modelo sobrestima o grupo masculino em precisão, fluência e prosódia, e o padrão persiste em todos os quatro grupos de características e na maioria dos regressores, indicando origem nos dados de treino mais do que no algoritmo. Paradoxalmente, o modelo é *mais preciso* (menor RMSE) para o grupo feminino em fluência e prosódia, apesar de esse grupo apresentar pontuações reais ligeiramente mais baixas em média. Segundo, há *degradação nas extremidades etárias*: o desempenho ótimo concentra-se em 14–16 anos, com sobrestimação de crianças de 6–10 e subestimação de adolescentes de 17–20. O grupo mais crítico é o masculino de 11–13 anos. Terceiro, *embeddings* baseados em *Gemini* introduzem um *viés específico* de acurácia favorável ao feminino, ausente nos modelos locais, sugerindo que o LLM multimodal captura sinais

linguísticos correlacionados com o gênero que extrapolam a informação acústica. Esses achados reforçam a necessidade de mitigar vieses ao implantar tais sistemas em contextos educacionais.

6. Considerações Finais

Este trabalho propôs um método para pontuar automaticamente a fluência de leitura, combinando o *encoder* de Whisper com o *embedding* multimodal do Gemini, com redução de dimensionalidade via PCA-256 e regressão por *ensembles* (*Stacking* e *Voting*). A configuração *Whisper* + *Gemini* + *PCA-256* foi avaliada contra linhas de base que incluem *Wav2Vec* + *BERT*, *Whisper* + *BERT* e *Gemini* isolado, em duas formulações da tarefa preditiva (classificação e regressão), sobre os quatro alvos do *SpeechOcean762*.

As evidências experimentais apontam que (i) o método proposto atinge PCC de até 0,77, superando outros trabalhos na base avaliada (QP1); e (ii) os regressores podem apresentar viés demográfico, sobrestimando sistematicamente o grupo masculino e degradando nas extremidades da distribuição etária, com padrão consistente entre representações e modelos, o que sugere origem nos dados de treino (QP2). A redução de dimensionalidade via PCA-256 mostrou-se uma estratégia eficaz para combinar Whisper e Gemini Embedding 2 sem incorrer em redundância, preservando ou melhorando o desempenho.

O estudo apresenta limitações que devem ser consideradas. A métrica de completude é praticamente constante no *corpus*, atuando como um alvo trivial e fornecendo pouca informação sobre a capacidade preditiva dos modelos. Não realizamos validação cruzada nem ajuste de hiperparâmetros, o que implica que as comparações relativas entre representações são robustas, mas os valores absolutos podem ser aprimorados. A análise de viés descreve correlações entre erros e atributos demográficos, mas não separa os efeitos confundidores decorrentes de faixa etária, do gênero e da proficiência de leitura. Por fim, o *corpus* cobre apenas a leitura em inglês por aprendizes não nativos, o que limita a generalização a outros idiomas e perfis de leitores.

Como trabalhos futuros, vislumbra-se uma linha promissora em torno da *destilação de conhecimento* [Hinton et al. 2015], motivada pelo principal achado deste estudo: a complementaridade entre o *encoder* acústico-fonético do Whisper e o *embedding* multimodal denso do Gemini Embedding 2 é o fator decisivo de desempenho, mas o acesso ao Gemini Embedding 2 ocorre via API paga, o que introduz latência e custos operacionais incompatíveis com cenários de implantação em escala ou em dispositivos de borda. Pretendemos, portanto: (a) aplicar *embedding distillation* [Reimers and Gurevych 2020] para treinar um *encoder* textual leve que reproduza as representações reduzidas por PCA-256 do Gemini Embedding 2, eliminando a dependência da API durante a inferência; e (b) explorar *cross-modal distillation* [Gou et al. 2021], em que um único modelo aluno baseado apenas em Whisper aprende a imitar a saída do regressor *Stacking* treinado sobre *Whisper* + *Gemini* + *PCA-256*, transferindo o sinal multimodal para um *pipeline* puramente acústico.

Declaração sobre uso de Inteligência Artificial

Reconhecemos o uso de IA Generativa para auxiliar na elaboração e revisão deste artigo, que foi devidamente revisado, e assumimos total responsabilidade pelo seu conteúdo.

Disponibilidade de artefatos

Os artefatos deste estudo serão disponibilizados pelos autores correspondentes.

Referências

- Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Bai, Y., Hubers, F., Cucchiari, C., and Strik, H. (2021). An asr-based reading tutor for practicing reading skills in the first grade. improving performance through threshold adjustment. ISCA.
- Bailly, G., Godde, E., Piat-Marchand, A.-L., and Bosse, M.-L. (2022). Automatic assessment of oral readings of young pupils. *Speech Communication*, 138:67–79.
- da Silva, G. C., Rodrigues, R. L., Amorim, A. N., Jeon, L., Albuquerque, E. X., Silva, V. C., da Silva, V. F., Pinheiro, A. L., Nunes, J. P., de Souza, S. X., et al. (2025). Assessing reading fluency in elementary grades: A machine learning approach. *Computers and Education: Artificial Intelligence*, 8:100411.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181.
- Gao, L., Tejedor-Garcia, C., Strik, H., and Cucchiari, C. (2024). Reading miscue detection in primary school through automatic speech recognition. *arXiv preprint arXiv:2406.07060*.
- Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.
- Harmsen, W., van Hout, R., Cucchiari, C., and Strik, H. (2025). Can asr generate valid measures of child reading fluency? In *Proc. Interspeech 2025*, pages 2395–2399.
- Hasbrouck, J. and Glaser, D. R. (2012). *Reading Fluency: Understanding and Teaching this Complex Skill (Training Manual)*. Gibson Hasbrouck & Associates.
- Hayati, D. S. and Saryanto, S. (2025). Development of deep learning methods to improve reading skills for elementary school students. *Jurnal Kajian Ilmu Pendidikan (JKIP)*, 6(3):1023–1033.
- Hinton, G. E., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.
- LaBerge, D. and Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, 6(2):293–323.

- Liu, J., Wumaier, A., Fan, C., and Guo, S. (2023a). Automatic fluency assessment method for spontaneous speech without reference text. *Electronics*, 12(8):1775.
- Liu, W., Fu, K., Tian, X., Shi, S., Li, W., Ma, Z., and Lee, T. (2023b). An asr-free fluency scoring approach with self-supervised learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Mohammadi, A., Koerich, A. L., Moro-Velazquez, L., and Cardinal, P. (2025). Automatic proficiency assessment in l2 english learners. *arXiv preprint arXiv:2505.02615*.
- Petscher, Y., O’Sullivan, J., Catts, H. W., Edwards, A., and Fitton, L. (2026). Evaluation of the consistency of a speech verification system with human raters in early literacy screening assessments. In *Frontiers in Education*, volume 11, page 1671946. Frontiers Media SA.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.
- van der Velde, M., Harmsen, W., Veldkamp, B. P., Feskens, R., Keuning, J., and Swart, N. (2025). Speech enabled reading fluency assessment: A validation study. *International journal of artificial intelligence in education*, 35(4):2569–2595.
- van der Velde, M., Molenaar, B., Veldkamp, B. P., Feskens, R., and Keuning, J. (2024). What do they say? assessment of oral reading fluency in early primary school children: A scoping review. *International journal of educational research*, 128:102444.
- Wang, Y., Wu, Z., Nese, J., Kamata, A., Nilabh, V., and Larson, E. C. (2025). A unified model for oral reading fluency and student prosody. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yıldız, M., Keskin, H. K., Oyucu, S., Hartman, D. K., Temur, M., and Aydoğmuş, M. (2025). Can artificial intelligence identify reading fluency and level? comparison of human and machine performance. *Reading & Writing Quarterly*, 41(1):66–83.
- Zhang, J., Zhang, Z., Wang, Y., Yan, Z., Song, Q., Huang, Y., Li, K., Povey, D., and Wang, Y. (2021). speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Proc. Interspeech 2021*.