Definição de heurística para identificação automática da fluência em leitura de crianças em fase de alfabetização

Cristiano Nascimento Silva[1], **André Luiz Vasconcelos Ferreira**[1], **Elias Cyrino de Assis**[1], **Jairo Francisco de Souza**[1,2]

[1]LApIC Research Group – UFJF – Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF) – Juiz de Fora, MG – Brasil

{cristiano.nascimento,andre.vasconcelos,elias.cyrino,jairo.souza}@ice.ufjf.br

Abstract. *Large-scale formative assessments provide data for planning and improving teaching methods and policies in education. These assessments, however, are very costly, generating demand for automating assessments for different areas of knowledge and academic years. This paper discusses the large-scale automatic assessment in the literacy process and aims to present a heuristic approach to minimize the error in the automatic identification of words read by students in reading fluency tests. Experiments were carried out with 6 experiments in 1039 audios of readings aloud by children in the first school years, in which the best experiment obtained an accuracy of 95,77%.*

Resumo. *Avaliações formativas em larga escala fornecem dados para o planejamento e melhoria nos métodos de ensino e políticas na educação. Essas avaliações, contudo, são muito custosas, gerando demanda para automatização de avaliações para diferentes áreas de conhecimento e anos letivos. Este trabalho discute a avaliação automática em larga escala no processo de alfabetização e tem como objetivo apresentar uma abordagem heurística para minimizar o erro na identificação automática de palavras lidas por alunos em testes de fluência. Foram realizados experimentos com 6 experimentos em 1039 áudios com leituras em voz alta de crianças dos primeiros anos escolares, no qual o melhor experimento obteve uma acurácia de 95,77%.*

1. Introdução

Avaliações formativas são testes em que se é possível acompanhar de forma permanente o processo de ensino-aprendizagem, permitindo assim, a adaptação das tarefas e métodos de aprendizagem por parte do professor [de Oliveira et al. 2007]. Avaliações formativas permitem verificar se os estudantes estão alcançando os objetivos esperados e a compatibilidade entre esses objetivos e os resultados atingidos [Oliveira 2002]. Diversos trabalhos abordam avaliações utilizadas em larga escala [Sousa 2014, Klinger 2008, Isac et al. 2015, Neumann et al. 2010, Kaplan and Huang 2021], mostrando sua ampla utilização em diferentes locais e contextos.

Tendo em vista a importância de avaliações em larga escala e considerando o seu custo em países de grandes dimensões como o Brasil, a automatização desse processo se torna relevante. Para isso, pode-se partir de soluções computacionais inteligentes na área

DOI: 10.5753/sbie.2022.224702

da Educação [Ceccon and Porto 2020, de Mira Gobbo et al. 2019]. A utilização de inteligência artificial (IA) se tornou intrínseco a diversas áreas, sendo a educação, uma delas [Al Braiki et al. 2020, Chassignol et al. 2018]. Em [Malik et al. 2019], por exemplo, são apresentadas diversas formas de utilização da IA em avaliações de aprendizagem. Além disso, a utilização de IA gera menor custo humano [Chassignol et al. 2018].

Em relação à avaliação da alfabetização, dificuldades como a avaliação de leitura, como custo de contrato de avaliadores humanos treinados [Carchedi et al. 2021] e demora na produção de resultados, a avaliação automática se mostra como uma alternativa adequada, mas que possui diversas dificuldades. Entre elas, pode-se citar a confiabilidade do sistema na identificação da pronúncia de palavras, principalmente considerando-se como público-alvo crianças em alfabetização, e a definição de métricas adequadas para determinar a qualidade da leitura.

Embora tecnologias para reconhecimento de fala tenham tido grandes avanços nos últimos anos em diversas áreas [Kabir et al. 2021], não é trivial o seu uso para avaliação de leitura em crianças, pois esta carrega características próprias, como uma maior frequência de pausas ao longo leitura, com pronúncias erradas, falso começos de palavras, repetições, entre outras [Proença et al. 2017]. O presente trabalho tem como objetivo apresentar uma abordagem heurística para minimizar o erro na identificação automática de palavras lidas. Para isso, foram feitos experimentos com 6 heurísticas em 1039 áudios selecionados aleatoriamente à partir de uma base de áudios de uma avaliação formativa de fluência em leitura de crianças do Ensino Fundamental I, mantendo um balanceamento entre leituras fluentes e não fluentes.

O trabalho se organiza da seguinte forma. Na Seção 3 é apresentada uma discussão sobre avaliações em larga escala e os trabalhos relacionados. Na seção 4 são apresentados os algoritmos para coleta de métricas de qualidade da leitura. Na seção 5 as métricas são utilizadas para avaliação da qualidade de leitura. Por fim, as Seções 6 e 7 apresentam os resultados dos experimentos realizados e as conclusões do trabalho, respectivamente.

2. Fundamentação teórica

Avaliações em larga escala tem se tornado mais populares ao longo dos anos com diversas iniciativas governamentais ao redor do mundo. Nesta seção, são apresentadas algumas dessas iniciativas, abrangendo avaliações em diversas áreas de conhecimento, mas dando maior atenção às avaliações sobre alfabetização. Por fim, as abordagens para identificação automática da qualidade de leitura são discutidas nos trabalhos relacionados.

2.1. Avaliação em larga escala no contexto mundial

No Brasil, avaliações em larga escala criadas pelo governo federal, como o SAEB, vêm sendo cada vez mais utilizadas como referência para o planejamentos de gestão [Sousa 2014]. Há uma tendência para a criação de mais avaliações parecidas com o SAEB por governos estaduais e municipais, em que os resultados obtidos da avaliação servem para realizar mudanças qualitativas nas escolas [Sousa 2014].

Em outros países avaliações em larga escala são utilizadas, para uso formativo ou não, como é o caso do Canadá [Klinger 2008], em que o uso destas se tornaram fundamentais para a definição de políticas e currículos educacionais. Em [Isac et al. 2015] é descrito as práticas de ensino nas escolas primárias e secundárias na Europa utilizando

dados de diversas avaliações em larga escala. São apresentadas quatro avaliações: para a educação primária são utilizados os dados de duas avaliações aplicadas na Europa em 2011, o TIMSS (*Trends in International Mathematics and Science Study*) e o PIRLS (*Program for International Reading Literacy Study*), e para a educação secundária, são utilizados os dados das avaliações PISA (*Program for International Student Assessment*) e TALIS (*Teaching and Learning International Study*), de 2012.

A Alemanha tradicionalmente não utilizava avaliações formativas em seu sistema educacional, até que uma participação no PISA mostrou que os estudantes alemães estavam em um nível menor que os demais países da Europa. Isto que causou uma reforma no sistema educacional alemão, mostrando o impacto que as avaliações formativas em larga escala têm para as políticas públicas [Neumann et al. 2010]. Já nos Estados Unidos é utilizado o NAEP (National Assessment of Educational Progress) como avaliação em larga escala desde 1970 em nível nacional e 1996 em nível estadual, o qual informações importantes para o monitoramento do rendimento acadêmico da população [Kaplan and Huang 2021]. Além disso, a participação de países nas avaliações internacionais como PIRLS, TIMSS e PISA aumentam anualmente [Dowd and Pisani 2013]. Apesar disso, países emergentes tendem a não participar dessas iniciativas devido ao alto custo da sua aplicação, o que tem incentivado a criação de avaliações voltadas para países mais pobres [Dowd and Pisani 2013], como é o caso da EGRA (Early Grade Reading Assessment), com objetivo de avaliar alfabetização.

O Brasil também possui iniciativas nacionais de avaliação na educação básica. A Provinha Brasil é utilizada para verificar o nível de alfabetização de crianças do segundo ano do ensino fundamental das escolas públicas brasileiras [Esteban and Wolf 2015]. De acordo com Dickel [Dickel 2016], a Avaliação Nacional da Alfabetização (ANA) tem sido usada como avaliação em larga escala com o objetivo de garantir a alfabetização de estudantes do ensino público nos primeiros 3 anos de escolaridade. Esses testes permitem um melhor planejamento do currículo do sistema educacional.

2.2. Trabalhos relacionados

Há alguns trabalhos que propõem soluções automáticas para avaliação automática de leitura. Em [Balogh et al. 2007, Black et al. 2007] são apresentadas propostas para avaliação automática da fluência em leitura em uma amostra de leituras de estudantes dos Estados Unidos. Em [Balogh et al. 2007], são utilizadas métricas sobre as palavras lidas, tempo de duração, quantidade de pausas e outras informações do áudio para avaliação da acurácia da leitura de adultos. Os resultados obtidos mostraram que o uso de avaliação automática é promissor. Em [Black et al. 2007], por sua vez, é realizada uma análise para identificar automaticamente erros de disfluência na leitura de 225 crianças até o segundo ano de escolaridade. As disfluências analisadas eram compostas de hesitações, sondagens, sussurros, paradas e questionamentos. Os autores criaram um método lexical que consistiu em adicionar um reconhecedor de fone restrito por palavra de destino à gramática do sistema de reconhecimento de fala (ASR) [Black et al. 2007]. O modelo utilizado foi o HMM (*Hidden Markov Model*) de 3 estados, com 16 misturas Gaussianas por estado. Ainda, em [Proença et al. 2017] é implementada uma ferramenta de predição automática da capacidade geral de leitura em voz alta avaliada com 284 crianças portuguesas. As disfluências mais comuns encontradas foram pausas intra-palavra, extensões fonéticas, falsos inícios, repetições e erros de pronúncia. Para a identificação das más

pronúncias, foi treinada uma rede neural voltada para reconhecimento de fonema, que foi utilizada para gerar uma matriz de probabilidades posteriores dos fonemas, para cada uma das frases do texto no tempo do áudio. Então, usando a razão de verossimilhança entre a palavra esperada e a dita, foi possível detectar más pronúncias.

Utilizando técnicas parecidas com os trabalhos anteriores, mas com um público-alvo distinto, Romana *et al* [Romana et al. 2021] objetivou identificar deficiências cognitivas em pessoas com Parkinson através da avaliação automática da leitura, utilizando uma base com 37 pessoas. Os autores usaram um sistema ASR com especialização do modelo (*finetune*) para aproximar da base testada. Os resultados obtidos, embora promissores, necessitam de novos estudos para verificar sua generalização em bases maiores. Para tentar generalizar a rede, em [Bailly et al. 2022] é proposto um *framework* para estimar classificações subjetivas multidimensionais do desempenho de leitura de jovens leitores. Foram usadas 1.063 leituras de 442 crianças, 84 leituras de 42 adultos e 6.853 avaliações subjetivas de 29 avaliadores humanos. Para o alinhamento automático da leitura, foi utilizado um modelo acústico GMM-HMM com três fones, um dicionário de pronúncia e modelos de trigramas para cada um dos textos lidos. Para alcançar os resultados, o processo definido pelos autores demanda diversos diversas etapas, o que demanda um custo considerável para preparação da base e das etapas num aplicação real.

Os trabalhos apresentados utilizaram diferentes abordagens, sendo a utilizada em [Proença et al. 2017] a mais similar ao feito no presente trabalho, mas com a diferença de que a matriz de probabilidades utilizada no presente trabalho é em relação aos caracteres na ordem esperada, ao invés dos fones, então a identificação automática fica atrelada ao alinhamento do que foi dito. O método utilizado neste trabalho também não depende de alterações na gramática, o que facilita o processo de reaproveitamento do modelo, trazendo vantagens em relação à sua presente em [Bailly et al. 2022]. Assim como em [Romana et al. 2021], o presente trabalho faz uso de um processo de *fine-tuning*, porém utilizando tecnologias mais atuais, como o *Wav2vec2*, o qual traz informações quanto às probabilidades de cada fone, permitindo uma gama maior de ações a partir dessas informações. Diferente de outros trabalhos, aqui é utilizada uma base com mais de mil leituras de crianças capturada de uma avaliação real de fluência em leitura, o que traz confiabilidade para os resultados do estudo.

3. Materiais e Métodos

Para determinar a qualidade da leitura infantil, o presente trabalho faz uso de duas métricas clássicas: QPL (quantidade de palavras lidas) e QPC (quantidade de palavras lidas corretamente). No contexto de avaliações automáticas, essas métricas são coletadas à partir de um áudio de uma leitura em voz alta. Todos os áudios são processados por um sistema de reconhecimento de fala, o qual gera uma transcrição do áudio. O sistema de reconhecimento de fala foi treinado utilizando uma grande quantidade de áudios, para identificar da forma mais adequada possível o que foi dito, abrangendo as variações por sotaque ou características regionais. A tarefa de *speech-to-text* pode gerar diversos erros de transcrição por conta da má qualidade de leitura ou por erros de identificação adequada do sistema, e o desafio de sistemas automáticos de avaliação da leitura é conseguir se utilizar de estratégias que melhor utilizem o áudio e/ou o transcrito para reconhecer as métricas de QPL e QPC mais próximas do esperado.

3.1. Algoritmos base para análise de transcritos

Este estudo utiliza-se de algoritmos para comparar o texto transcrito com o texto de referência, ou seja, o que se espera que a criança tenha lido. Para isso, são utilizados como base duas abordagens de comparação: um algoritmo de busca aproximada (*fuzzy text search* ou $f_{search}(t)$) e um algoritmo de alinhamento de texto (*text alignment* ou $f_{align}(t)$).

O $f_{search}(t)$ consiste em uma técnica para encontrar padrões em textos. O diferencial do método está no fato dele buscar palavras aproximadas. Por exemplo, se a palavra "sapo" for buscada no texto, o algoritmo irá comparar cada parte do texto e ver o quanto elas se aproximam do termo objetivo permitindo um conjunto determinado de erros. Assim, palavras como "sapa", "saco" ou "sapinho" podem ser identificadas. No contexto deste estudo, a técnica se mostra interessante pois as transcrições das leituras das crianças muitas vezes possuem padrões de erros que se aproximam da redação original. Isso acontece por diversos motivos, como quando uma criança possui apresenta lentidão na leitura e o sistema de reconhecimento transcreve palavras de forma condizente com a pronúncia mas erroneamente, como "palhaço" e "palha aço". Assim, pode-se utilizar esse algoritmo para buscar uma sequência de termos dentro de uma janela do texto transcrito.

O $f_{align}(t)$ compara o todo o transcrito com o texto de referência, realizando um alinhamento entre os dois. Alinhar duas sequências de palavras significa identificar quais operações de remoção, substituição e inclusão de palavras são necessárias para fazer o texto transcrito se transformar no texto de referência. Com isso, é retornada a quantidade de operações e de correspondências corretas mínimas para realizar o alinhamento. Por exemplo, o alinhamento de "a carroça sem freio pode correr apressada" e "o carro sim sem freio corre apressa apressado" gera um alinhamento entre "sem", "freio", "apressado", com quatro substituições (a $\rightarrow$ o, carroça $\rightarrow$ carro, correr $\rightarrow$ corre), duas inserções (sim, apressa) e uma remoção (pode). Estes dois algoritmos, $f_{search}(t)$ e $f_{align}(t)$, são utilizados como base para um conjunto de estratégias para cálculo das métricas de QPL e QPC.

3.2. Heurísticas para cálculo do QPC

Foram definidas duas heurísticas para cálculo do QPC. A heurística HC_1 consiste na utilização do número de correspondências corretas resultantes do alinhamento feito pelo $f_{align}(t)$. Assim, utiliza-se exclusivamente o texto transcrito.

Por sua vez, a heurística HC_2 é calculada utilizando a matriz de probabilidades de fones para cada *frame* do áudio do *posteriorgrama*. O posteriorgrama é a matriz de saída da rede neural que processa o áudio e realiza o reconhecimento de fala. Neste estudo, utiliza-se um *fine-tuning* de uma rede Wav2Vec2 com áudios em português. Um algoritmo de alinhamento de áudio, então, compara o que é pronunciado no áudio com os caracteres esperados para uma determinada faixa de tempo e atribui um índice de confiança S para a leitura de cada palavra. Seja s_i a porcentagem do fone i no posteriorgrama e t_i o número de *frames* do áudio em que o fone i foi pronunciado, isto é, sua duração, então a Equação 1 é utilizada para calcular o índice de confiança para cada palavra esperada.

$$S = \frac{\sum_{i=0}^{k} s_i t_i}{\sum_{i=0}^{k} t_i} \tag{1}$$

O QPC é calculado como a quantidade de palavras que possuem um índice de confiança maior que um limiar previamente definido. Palavras com confiança abaixo do

limiar são consideradas de leitura incorreta. Palavras curtas, como artigos e preposições, tendem a ser ignoradas pelo avaliador humano, muitas vezes por conta da velocidade da leitura, qualidade do áudio ou desatenção do avaliador. Como espera-se encontrar uma solução que mimetize o comportamento do avaliador humano, é importante que a abordagem não seja rígida demais. Assim, quando uma palavra muito curta é considerada de leitura incorreta por HC_2, estas são contadas como *palavras suspeitas* ou QPS. A quantidade de palavras de leituras suspeitas pode ser determinante para verificar a confiabilidade do QPC gerado pela heurística, o que será provado na análise dos resultados finais.

3.3. Heurísticas para cálculo do QPL

Foram definidas 6 heurísticas para cálculo do QPL, todas criadas a partir da utilização de informações retiradas dos algoritmos $f_{align}(t)$ ou $f_{search}(t)$, que se dispõem como:

- HL_1: utiliza o QPC calculado em HC_1 somado à quantidade de substituições presentes no $f_{align}(t)$.
- HL_2: valor de HL_1 somado ao número de deleções presentes no $f_{align}(t)$.
- HL_3: similar à HL_2 mas soma apenas deleções que aparecem em sequências de no máximo 3.
- HL_4: similar à HL_1 mas soma apenas deleções até que apareça uma sequência de deleções maior que 3.
- HL_5: utiliza a posição da última palavra marcada como correta no retorno do $f_{align}(t)$.
- HL_6: utiliza o $f_{search}(t)$ para encontrar na referência a posição dos últimas 10 palavras do transcrito permitindo um erro de até 10 caracteres. Caso não tenha sido encontrado uma sequência válida, utiliza o valor de $f_{align}(t)$. Diferente das heurísticas anteriores, é aplicado uma regra para confirmação do valor de QPC e QPL nesta heurística. Caso a razão entre o QPS e QPC seja maior que um limiar de descarte, os valores de QPL e QPC são zerados, ou seja, a criança tem um nível de erros tão alto que o resultado das métricas para esse áudio não é confiável.

3.4. Base de dados

Foram utilizados áudios obtidos através de avaliações de leitura realizadas pela Fundação CAEd/UFJF, que consistem de leituras de voz alta de crianças do Ensino Fundamental I, as quais foram instruídas a ler um texto narrativo composto de 250 palavras. Como é comum em outros estudos, os áudios contém apenas o primeiro minuto de leitura da criança, o que melhor representa o comportamento da criança em uma leitura à primeira vista. Por conta do número de experimentos e do tempo para execução de todos eles, foi selecionada uma amostra de 1.039 áudios selecionados aleatoriamente de um conjunto de 55.000 áudios. A amostra foi definida para representar a base original com 95% de grau de confiança e 3% de margem de erro. Todos os áudios dessa base foram avaliados manualmente por corretores especialistas em alfabetização. Para cada áudio, os corretores informaram a última palavra do texto lido pelo falante, considerado aqui como QPL, e a quantidade de palavras lidas corretamente (QPC) de acordo com o texto de referência, ocultando informações pessoais das crianças leitoras.

A base se divide entre leitores fluentes (521 áudios) e disfluentes (518 áudios). A definição das classes é determinada na base segundo estudos da área de avaliação em língua portuguesa, a qual determina que, para essas condições de avaliação, uma leitura

esperada para alunos nesta faixa etária é de QPC maior que 65 com 90% de precisão na leitura, isto é, a razão entre QPC e QPL.

3.5. Definição dos experimentos

Foram realizados experimentos com todas as heurísticas. Como implementação dos algoritmos de $f_{search}(t)$ e $f_{align}(t)$, foram utilizados algoritmos disponíveis gratuitamente. Para busca aproximada foi utilizada a solução X[1] e para alinhamento foi utilizado o Sclite[2]. Ainda, foi definido como limiar para confiança do QPC um valor de 0,125 e limiar de descarte para a métrica HL_6 de 0,3. Os limiares e os valores adotados no cálculo do HL_4 (grupos de mais de 3 remoções) e HL_6 (janelas de 10 palavras e erro de 10 caracteres) foram definidos à partir de experimentos prévios, os quais foram omitidos desse artigo por questão de espaço.

4. Análise dos resultados

Uma vez calculado o valor de QPL e QPC para cada heurística, os resultados foram comparados com as avaliações humanas e, assim, calculados os erros de QPL e de QPC. A Tabela 1 apresenta o erro QPL de cada heurística em relação à média absoluta ($\bar{X}_{QPL}$) e desvio padrão (σ).

Tabela 1. Resultados médios de erro QPL e QPC dos experimentos.

Heurística	$\bar{X}_{QPL}$	σ
HL_1	8,71	14,12
HL_2	9,95	15,30
HL_3	8,48	14,38
HL_4	8,26	13,84
HL_5	36,16	45,33
HL_6	**4,18**	11,20

Para o cálculo da média do erro QPL absoluto, a heurística HL_6 apresentou resultado muito superior às demais soluções. Em seguida, as melhores heurísticas foram HL_1, HL_3 e HL_4. HL_2 teve um resultado melhor que HL_1, apesar de pouca diferença e com maior desvio padrão. Em relação ao QPC, as heurísticas HL_1 a HL_5 compartilham o mesmo cálculo. O erro médio absoluto do QPC foi igual a 8,22 ($\sigma = 12,81$). Por sua vez, o erro médio absoluto da heurística HL_6 foi igual a 4,36 ($\sigma = 10,65$), ou seja, quase metade do erro das outras métricas.

4.1. Impacto das métricas para classificação de fluência

Para melhor entender o impacto das métricas encontradas, elas foram submetidas à uma tarefa de classificação. Em testes de fluência de leitura, é comum adotar como fluentes os estudantes que leram mais de 65 palavras por minuto com 90% de precisão na leitura. No contexto desse trabalho, a base de dados adotada possui áudios de 60 segundos. Assim, considera-se como fluente as leituras com QPC $>$ 65 e precisão como sendo a razão entre QPC e QPL. Uma vez classificados todas leituras, elas foram comparadas com as classificações dos avaliadores humano, gerando as métricas apresentadas na Tabela 2.

[1] `https://pypi.org/project/fuzzysearch/`
[2] `https://github.com/usnistgov/SCTK`

Tabela 2. Resultados em relação à classificação em fluentes e não fluentes dos experimentos.

Heurística	Precisão Fluentes	Revocação Fluentes	Precisão Não Fluentes	Revocação Não Fluentes	Acurácia
HL_1	99,43%	67,37%	75,22%	99,61%	**83,45%**
HL_2	99,64%	53,36%	68,03%	99,81%	76,52%
HL_3	99,67%	57,39%	69,96%	99,81%	78,54%
HL_4	99,34%	58,16%	70,30%	99,61%	78,83%
HL_5	100,00%	22,07%	56,06%	100,00%	60,92%
HL_6	98,97%	92,51%	92,93%	99,03%	**95,77%**

A precisão expressa quantas classificações como fluente estavam certas, dentre todas as classificações como fluentes calculadas a partir do reconhecimento automático. A revocação mostra a proporção da quantidade de classificações como fluente, em comparação com a quantidade de classificações como fluente de acordo com a avaliação manual. Por fim, a acurácia indica a porcentagem de acerto geral do reconhecimento, seja tanto para classe de fluentes como para não fluentes. As melhores classificações foram as realizadas utilizando as métricas geradas por HL_6 e HL_1. Verifica-se, que as métricas da heurística (HL_6 trazem resultados significativamente maiores que HL_1 (95,77% vs 83,45%) de acurácia. Embora o erro absoluto médio do QPL e QPC do HL_6 é quase metade do HL_1, a acurácia não se compara no mesmo nível. Isso se dá porque erros de QPC, por exemplo, acima do limiar de 65 palavras podem continuar situando a leitura como fluente, caso se mantenha a razão entre QPL e QPC.

4.2. Análise dos erros e abordagem em duas etapas

Embora os resultados da classificação automática com as métricas geradas pelas heurísticas tenha encontrado valores satisfatórios, pode-se encontrar resultados mais confiáveis ao se adotar uma abordagem que contemple uma segunda fase com uma avaliação manual de um conjunto de áudios com alta probabilidade de terem sido mal classificados automaticamente. Toda solução de predição automática está sujeita a erros e o uso de uma segunda etapa de classificação é comum em aplicações comerciais quando deseja-se minimizar o erro da máquina, mesmo que aumentando o custo com trabalho humano. O desafio, contudo, é alcançar um modelo de classificação que gere um alto volume de classificações corretas e que consiga selecionar um conjunto idealmente pequeno de instâncias que necessitam de verificação humana.

Neste sentido, considerando que uma abordagem para avaliação em larga escala na área da educação que utiliza necessita minimizar os erros de classificação, foi feita uma análise dos falsos positivos e falsos negativos das classificações para determinar uma regra que permita selecionar leituras com grande chance de erro na classificação. Como a classificação de fluência se dá pelo limiar de QPC e precisão de leitura, foram testados diferentes faixas de valores até encontrar uma faixa que minimize o conjunto filtrado para avaliação manual e maximize a acurácia do conjunto final.

O filtro adotado consiste em separar os áudios com QPC entre 55 e 66 ou QPC acima de 40 com precisão entre 0,5 e 0,9. Com o filtro adotado, percebe-se que a abordagem automática possui uma avaliação mais rígida que o avaliador humano, ou seja, tende a atribuir valores mais baixos de QPC e fluência do que o atribuído pelo avaliador humano. Leituras com valores pertencentes a uma faixa próxima ao limiar das classes estão mais sujeitas a serem atribuídos à classe errada e, neste estudo, a faixa possui mais

extensão para valores menores que 65, o que influencia também na faixa de precisão. A Tabela 3 apresenta os resultados utilizando o filtro nas duas heurísticas com melhores resultados de classificação (Tabela 2).

Tabela 3. Resultado da classificação após aplicação do filtro

Heurística	CLASSIFICAÇÃO					MATRIZ				
	Precisão Fluentes	Revocação Fluentes	Precisão Não Fluentes	Revocação Não Fluentes	Acurácia	TP	FP	TN	FN	FILTRO
HL_1	99,41%	97,41%	98,25%	99,61%	98,71%	338	2	506	9	194
HL_6	98.97%	98.37%	98.41%	99.00%	**98.69%**	482	5	496	8	**48**

O valor de acurácia entre as abordagens possui alteração significativa após uso do filtro em comparação ao resultado anterior: 98,71% vs 83,45% para HL_1 e 98,69% vs 95,77% para HL_6. A aplicação do filtro, por outro lado, fez com que essas duas heurísticas tivessem acurácia praticamente idêntica. Contudo, percebe-se que o número de áudios filtrados em HL_6 abordagem é quase 75% menor que em HL_1 (48 vs 194). Menos áudios filtrados significa menos áudios a serem enviados para a correção manual e, neste caso, diminuiu-se de 18,76% dessa base (a qual foi artificialmente equilibrada entre classes) para 4,6%. Numa projeção para a base completa de áudios, onde 5,39% dos áudios foram classificados como fluentes pelos corretores manuais, espera-se que apenas 0,9% dos áudios precisem de avaliação manual. Isso significa que a correção manual de pouco mais de 5000 áudios filtrados da base completa permitiria subir a taxa de acerto do sistema de 95,77% para 98,69%.

5. Conclusão

O trabalho realizado teve como objetivo melhorar o cálculo do QPL e QPC a partir da utilização de diferentes abordagens. Verificou-se a diferença substancial entre uma das abordagens em relação às demais na geração das métricas. Como esperado, provou-se que quanto menor o erro na geração das métricas, mais acurado será o resultado em uma tarefa, como a tarefa de classificação. Porém, vale destacar que os resultados mostram que a diferença na qualidade dos resultados não mantém a mesma proporção da diferença do erro de geração das métricas.

Esse trabalho apresentou importantes contribuições para a pesquisa e prática em avaliação de fluência, onde destaca-se: (i) um conjunto consistente de heurísticas que podem ser utilizadas para geração das métricas mais usadas por trabalhos na área de avaliação de fluência; (ii) um estudo do impacto dessas heurísticas no contexto da avaliação automática de fluência em língua portuguesa, a qual é uma área ainda carente desse tipo de trabalho; (iii) resultados utilizando uma amostra selecionada aleatoriamente de uma base de 55.000 áudios de uma teste real de fluência em leitura, definida para representar a base original com 95% de grau de confiança e 3% de margem de erro nos resultados gerados; (iv) um estudo de impacto do uso de uma abordagem de classificação em duas etapas, a qual permite balancear entre o custo de uma avaliação manual para as instâncias com maior chance de erro de classificação pela máquina e qualidade do resultado final.

Não existe uma investigação que seja definitiva e os resultados apresentados neste trabalho não podem ser considerados conclusivos. Embora tenha sido utilizada uma

população volumosa (55000 leituras em voz alta de crianças do Ensino Fundamental I do estado do Maranhão), estudos com outras populações são necessários para se definir a melhor heurística. Estes novos estudos podem abranger populações de diferentes estados brasileiros, mas também diferentes itens narrativos e anos escolares. Ainda, a abordagem apresentada é independente de idioma, embora suas etapas precisam ser adaptadas para outros idiomas, o que pode impactar no resultado das heurísticas. A qualidade das métricas de QPL e QPC são fortemente dependentes da qualidade do modelo acústico utilizado. Novos estudos precisam ser realizados para verificar o impacto de diferentes tipos de modelos acústicos, uma vez que algumas aplicações avaliativas podem apresentar requisitos que não se adequam bem para alguns modelos, como restrições de tamanho do modelo, tempo de uso de GPU, *throughput* e geração das métricas em *streamings* de áudios.

Por fim, ressalta-se que os resultados visam verificar a capacidade das soluções em imitar o comportamento do avaliador humano. A tarefa de avaliação de áudios de leituras está sujeita a alguns efeitos que podem interferir na nota do avaliador, como uma subjetividade na avaliação de algumas palavras, viés de confirmação, cansaço do avaliador e discordância de julgamento entre avaliadores. A base utilizada não pode ser considerada um padrão ouro e, assim, está sujeita a erros de avaliação. Por consequência, é necessário verificar as discordâncias entre a avaliação automática e humana. O padrão de erro nas métricas de QPL e QPC nos experimentos realizados mostrou um comportamento mais rígido da avaliação da máquina do que do avaliador humano, o que pode demonstrar que o avaliador humano tende a considerar como corretas certas variações de leitura de palavras, como troca ou omissão de fones.

Referências

Al Braiki, B., Harous, S., Zaki, N., and Alnajjar, F. (2020). Artificial intelligence in education and assessment methods. *Bulletin of Electrical Engineering and Informatics*, 9(5):1998–2007.

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.

Balogh, J., Bernstein, J., Cheng, J., and Townshend, B. (2007). Automatic evaluation of reading accuracy: assessing machine scores. In *Workshop on Speech and Language Technology in Education*.

Black, M., Tepperman, J., Lee, S., Price, P., and Narayanan, S. S. (2007). Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment. In *Eighth Annual Conference of the International Speech Communication Association*.

Carchedi, L. C., Barrére, E., and de Souza, J. F. (2021). Avalia online: um sistema para avaliação em larga escala de testes de fluência de leitura. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 01–11. SBC.

Ceccon, D. and Porto, J. (2020). Bcs: Jogos digitais no auxílio do desenvolvimento de crianças especiais com atraso na linguagem. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 522–531, Porto Alegre, RS, Brasil. SBC.

Chassignol, M., Khoroshavin, A., Klimova, A., and Bilyatdinova, A. (2018). Artificial intelligence trends in education: a narrative overview. *Procedia Computer Science*, 136:16–24.

de Mira Gobbo, M. R., Barbosa, C. R., Morandini, M., and Mafort, F. (2019). Aplicativo para ganho de vocabulário e auxílio na alfabetização destinado às crianças com transtorno do espectro autista. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1111.

de Oliveira, E. d. S. G., Cunha, V. L., da Encarnação, A. P., Santos, L., de Oliveira, R. A., and da Silva Nunes, R. (2007). Uma experiência de avaliação da aprendizagem na educação a distância. o diálogo entre avaliação somativa e formativa. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, 5(2):39–55.

Dickel, A. (2016). A avaliação nacional da alfabetização no contexto do sistema de avaliação da educação básica e do pacto nacional pela alfabetização na idade certa: responsabilização e controle. *Cadernos Cedes*, 36:193–206.

Dowd, A. J. and Pisani, L. (2013). Two wheels are better than one: the importance of capturing the home literacy environment in large-scale assessments of reading. *Research in Comparative and International Education*, 8(3):359–372.

Esteban, M. T. and Wolf, C. C. (2015). Um olhar para a alfabetização a partir dos exames nacionais. *Revista de estudios e Investigación en Psicologia y Educación*, pages 160–164.

Isac, M. M., da Costa, P. D., Araújo, L., Calvo, E. S., and Albergaria-Almeida, P. (2015). Teaching practices in primary and secondary schools in europe: Insights from large-scale assessments in education. *JRC Science and Policy Report*.

Kabir, M. M., Mridha, M., Shin, J., Jahan, I., and Ohi, A. Q. (2021). A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. *IEEE Access*.

Kaplan, D. and Huang, M. (2021). Bayesian probabilistic forecasting with large-scale educational trend data: a case study using naep. *Large-scale Assessments in Education*, 9(1):1–31.

Klinger, D. (2008). The evolving culture of large-scale assessments in canadian education. *Canadian Journal of Educational Administration and Policy*, (76).

Malik, G., Tayal, D. K., and Vij, S. (2019). An analysis of the role of artificial intelligence in education and teaching. In *Recent Findings in Intelligent Computing Techniques*, pages 407–417. Springer.

Neumann, K., Fischer, H. E., and Kauertz, A. (2010). From pisa to educational standards: The impact of large-scale assessments on science education in germany. *International Journal of Science and Mathematics Education*, 8(3):545–563.

Oliveira, G. d. (2002). Avaliação formativa nos cursos superiores: verificações qualitativas no processo de ensino-aprendizagem e a autonomia dos educandos. *OEI-Revista Iberoamericana de Educación. Disponível em:. Acesso em*, 15.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., and Perdigão, F. (2017). Automatic evaluation of reading aloud performance in children. *Speech Communication*, 94:1–14.

Romana, A., Bandon, J., Perez, M., Gutierrez, S., Richter, R., Roberts, A., and Provost, E. M. (2021). Automatically detecting errors and disfluencies in read speech to predict cognitive impairment in people with parkinson's disease. In *Interspeech*, pages 1907–1911.

Sousa, S. Z. (2014). Concepções de qualidade da educação básica forjadas por meio de avaliações em larga escala. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 19:407–420.