Classificação automática de áudios de leituras de pseudopalavras para avaliação em larga escala de fluência da leitura de crianças em fase de alfabetização

Elias Cyrino de Assis[1], **André Luiz Vasconcelos Ferreira**[1], **Cristiano Nascimento Silva**[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. *The pseudoword reading test is used in several large-scale assessments that seek to assess the reading fluency of children in the literacy phase. As with the other items that make up the fluency assessments, assessing the pseudoword test is a costly task, from which the development of ASR systems to automate the assessment process arises. In this work, an approach to automatically evaluate pseudoword readings using a pre-trained self-supervised model is presented. Three experiments were carried out with different strategies to calculate reading metrics. The performance of the strategies was compared with the evaluation of the human corrector.*

Resumo. *O teste de leitura de pseudopalavras é utilizado em diversas avaliações em larga escala que buscam avaliar a fluência em leitura de crianças em fase de alfabetização. Assim como os demais itens que compõem as avaliações de fluência, avaliar o teste de pseudopalavras é uma tarefa custosa, de onde surge o desenvolvimento de sistemas de ASR para automatizar o processo de avaliação. Neste trabalho é apresentada uma abordagem para avaliar automaticamente leituras de pseudopalavras utilizando um modelo autosupervisionado pré-treinado. Três experimentos foram realizados com diferentes estratégias para cálculo das métricas de leitura. O desempenho das estratégias foi comparado com a avaliação do corretor humano.*

1. Introdução

A capacidade de uma leitura fluente é um fator de alta importância para que o indivíduo seja inserido e participe ativamente da sociedade [Dias et al. 2016, Farrall and Ashby 2019]. Muitos esforços têm sido feitos para implementar políticas para a correta apreensão da língua materna, onde a avaliação de fluência em leitura oral tem se destacado como um dos pilares para o desenvolvimento da alfabetização [National Reading Panel 2000]. Um dos aspectos da leitura oral avaliados é a capacidade de decodificação dos símbolos da língua materna, que deve ocorrer de forma automática, sem ou com mínimo esforço, e com precisão, respeitando a pronúncia correta [Rasinski 2004, Batista 2011]. Dentre os itens utilizados na aplicação de testes de

DOI: 10.5753/sbie.2022.224701

fluência em leitura oral se encontra o teste de leitura de pseudopalavras. A leitura de pseudopalavras fornece uma análise distinta e objetiva da consciência fonética e das correspondências entre letras e sons independente de familiaridade ou contexto [Proença 2018] e pode contribuir para o estímulo em áreas do cérebro relacionadas à recuperação lexical [Mechelli et al. 2003]. Evidências psicométricas apontaram o teste de pseudopalavras como um dos instrumentos padrão-ouro para avaliação da decodificação grafêmica [Pinheiro and de Araújo Vilhena 2022], justificando a presença deste item no processo de avaliação da leitura oral para a composição de um sistema de avaliação mais completo.

Com os investimentos feitos nas últimas décadas para que cada vez mais crianças ao redor do mundo tivessem acesso à educação, a preocupação com a qualidade do ensino também aumentou. Compreender os dados gerados a partir das avaliações e métodos empregados tanto a nível nacional como internacional é crucial para que comparações possam ser feitas e intervenções adequadas em determinadas áreas do ensino venham a ser aplicadas [Cresswell et al. 2015, Wagemaker 2014]. Contudo, o processo de avaliação da educação em larga escala é altamente custoso, sobretudo para a avaliação de fluência em leitura oral. Muito recurso financeiro, humano e tempo é necessário para que as avaliações sejam feitas de forma adequada. Além de gerar uma grande quantidade de material a ser avaliado, a própria ação de aferir os itens avaliativos é exaustiva, requerendo muito esforço cognitivo. O desenvolvimento de tecnologias computacionais para atender à avaliação em larga escala pode ajudar significativamente na redução dos custos relacionados [Almeida Silva et al. 2021]. Dentro do contexto de avaliação de fluência em leitura oral, a tecnologia que mais tem ganhado destaque é o uso de técnicas para reconhecimento automático de fala (ASR - *Automatic Speech Recognition*).

Embora a tecnologia de ASR seja bem estabelecida em domínios mais gerais, alguns fatores dificultam sua aplicação na avaliação de leitura oral em crianças, principalmente para o teste de pseudopalavras. Como se tratam de palavras não existentes na língua, não existem bases de dados para treinar um sistema de reconhecimento automático específico para esse fim. A confecção de uma base com as informações necessárias para treinar um sistema de ASR nesse contexto gera custos altos. Ainda, tem-se que as características acústicas da fala de crianças em fase de alfabetização diferem das características presentes na fala de adultos [Proença 2018], o que dificulta o uso de sistemas de ASR treinados com leituras de adultos para avaliar a leitura de crianças [Elenius and Blomberg 2004, Shivakumar and Georgiou 2020]. Além disso, a leitura de crianças costuma ser mais pausada, sendo bem frequente a ocorrência de silabações, repetições e falsos começos. Tais desvios do que seria considerado uma leitura fluida impõem desafios ao processo de alinhamento da transcrição com o áudio, dificultando a avaliação se a pronúncia de determinada palavra foi ou não adequada. A qualidade do áudio e o som ambiente são outros fatores que influenciam no desempenho do sistema, uma vez que a presença de ruídos e interrupções, como a própria qualidade do equipamento de captação do áudio, pode atrapalhar o reconhecimento de forma significativa. Por fim, deve-se também considerar o público-alvo dessas avaliações. Crianças nos primeiros anos letivos podem apresentar um comportamento que diverge do protocolo de aplicação do teste, resultando em áudios com falas inesperadas ou leituras em ordem imprevista.

O presente trabalho tem como objetivo apresentar uma nova abordagem para avaliação de leituras de pseudopalavras em língua portuguesa e avaliar o desempenho

de um modelo pré-treinado para aprender representações da fala em áudios não rotulados. Foram usados 1560 áudios de leituras reais de pseudopalavras do estado do Espírito Santo referentes a uma avaliação de fluência em larga escala aplicada em alunos do 2º ano do ensino fundamental. O resultado do sistema foi comparado com a nota atribuída por especialistas treinados para esta avaliação.

O trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados à avaliação automática de leitura de pseudopalavra; a Seção 3 discute os critérios utilizados no teste de pseudopalavras e apresenta considerações sobre o modelo utilizado para reconhecimento dos áudios; a Seção 4 descreve como foi feito o reconhecimento automático para pseudopalavras e o cálculo das métricas de leitura; a Seção 5 apresenta os experimentos realizados e os resultados obtidos após a comparação das notas atribuídas pela abordagem automática e pelo corretor humano; por fim, a Seção 6 apresenta conclusões sobre os resultados obtidos e discute os próximos passos.

2. Trabalhos relacionados

Alguns trabalhos apresentam ou avaliam abordagens automáticas de avaliação de fluência em leitura oral para o teste de pseudopalavras. Em [Duchateau et al. 2007] os autores utilizam um sistema de reconhecimento automático para a língua alemã, treinado com áudios de leituras de sentenças com crianças de 5 a 11 anos, composto por duas camadas. A primeira camada, *phoneme recognizer*, é responsável por gerar um reconhecimento a nível de fonemas, utilizando para isso uma representação chamada de *lattice* de fonemas, onde é possível encontrar o caminho mais provável para um áudio de entrada que corresponde aos fonemas que se deseja reconhecer. A segunda camada, *lattice search module*, executa uma busca no *lattice* de fonemas gerado e retorna a representação do reconhecimento em nível de palavra. Um estágio de pós-processamento é aplicado para identificar se a representação recuperada indica ou não uma leitura correta, utilizando medidas de confiança acústica. A pontuação final da leitura da criança é calculada como a razão entre o tempo gasto na leitura de 40 palavras ou pseudopalavras e o número de palavras ou pseudopalavras lidas corretamente, retornando uma medida do tempo médio por palavra ou pseudopalavra correta. A leitura é então classificada em um de cinco grupos de classificação, a depender da pontuação calculada. Em [Yılmaz et al. 2014], os autores estendem o trabalho de [Duchateau et al. 2007] aplicando um esquema de decodificação mais flexível onde substituições, deleções e inserções de fonemas são permitidas e estende o *lattice* de fonemas da primeira camada utilizando uma matriz de confusão de fonemas que modela as confusões de fonemas típicas em uma língua.

Por sua vez, em [Proença et al. 2017a] os autores extraem diversas características acústicas do áudio e aplicam um processo de anotação automática onde ocorrências de leituras inadequadas, falsos começos, repetições, pausas entre palavras e extensões de palavras são capturadas. Com as características extraídas e possíveis erros de pronúncia anotados, sistemas de regressão gaussiana para a língua portuguesa foram treinados e avaliados, sendo o desempenho do sistema aferido comparando-se a sua classificação com a avaliação manual. A base de dados utilizada no estudo contém leituras de sentenças e pseudopalavras na língua portuguesa, onde os falantes possuem a faixa etária de 6 a 10 anos. As leituras foram classificadas baseando-se em uma pontuação de 0 a 5 dada pelos avaliadores manuais. De maneira semelhante a [Proença et al. 2017a] e [Yılmaz et al. 2014], em [Proença et al. 2017b] é apresentada uma abordagem que se uti-

liza de *lattices* que possibilitam a identificação de repetições de palavras e falsos começos. Também são utilizadas múltiplas características do áudio derivadas da medida GoP (*Goodness Of Pronunciation*) e extraídas com o auxílio da decodificação de fonemas e do alinhamento forçado. Estas características são usadas para treinar classificadores como redes neurais, máquinas de vetor suporte e regressão linear.

O uso de jogos educacionais para alfabetização é outra frente de trabalho que tem ganhado destaque [Passos et al. 2019] e aplicações têm sido desenvolvidas para conter os atrasos no aprendizado da língua até mesmo em crianças especiais [Ceccon and Porto 2020, de Mira Gobbo et al. 2019, Pantoja et al. 2018]. Assim, também têm surgido soluções para avaliação de leitura de palavras ou pseudopalavras no contexto de gamificação ao embutir sistemas de ASR em jogos educativos para que o aprendizado da criança ocorra de maneira lúdica. [Silva et al. 2019] apresenta uma abordagem computacional gamificada que auxilia professores no diagnóstico da capacidade de decodificação das palavras nos anos iniciais do ensino fundamental. Um jogo denominado SpaceGEMS é desenvolvido para a coleta lúdica do material e utiliza ASR para avaliar a fluência das leituras e técnicas de gamificação para engajar os alunos durante a atividade. [Hautala et al. 2020] avalia a confiabilidade de um sistema de CGBA (*computarized game-based assesment*) para identificar leitores com dificuldade de leitura. Um estudo em larga escala é realizado com estudantes finlandeses da primeira a quarta série. O objetivo foi utilizar o GBA para identificar leitores que possuem desempenho abaixo de uma média dada por testes tradicionais feitos manualmente. Entre os itens avaliados estão a leitura de testes de palavras e pseudopalavras. As leituras são gravadas e processadas por um sistema de ASR. No teste de pseudopalavras o leitor deve ler, da maneira mais rápida e acurada possível, 30 pseudopalavras. O número de pseudopalavras pronunciadas e de pseudopalavras lidas corretamente em um minuto foram utilizadas como métricas de qualidade da leitura.

A abordagem apresentada e avaliada no presente trabalho se difere das demais pelo uso de um modelo auto-supervisionado pré-treinado para aprendizado de padrões de fala e que foi adaptado para realizar o reconhecimento automático de pseudopalavras em língua portuguesa. Novas estratégias para determinar a corretude das leituras bem como quais pseudopalavras foram lidas também foram desenvolvidas. A vantagem do uso de uma abordagem auto-supervisionada é a capacidade de alcançar bons resultados com conjuntos de treinamento reduzidos e, por consequência, pode permitir a popularização da tecnologia para diversas aplicações na área de Informática na Educação.

3. Fundamentos conceituais

As pseudopalavras são uma sequência de grafemas sem significado e construídas respeitando-se as estruturas das palavras aceitas em uma língua [Salles and Parente 2007]. Testes de leitura de pseudopalavras objetivam avaliar a capacidade de decodificação da criança e contém um conjunto de pseudopalavras criado segundo critérios definidos de acordo com o ano letivo do público-alvo. O teste aplicado pela Fundação CAEd, a qual forneceu os dados para os experimentos realizados no presente trabalho, é composto por um quadro de pseudopalavras onde a leitura deve ocorrer sequencialmente da esquerda para a direita. O leitor é instruído a ler as pseudopalavras na ordem estabelecida, sendo requisitada a leitura do máximo de entradas possível dentro do tempo de um minuto. A leitura é gravada em áudio para análise posterior.

3.1. Critérios para avaliação de leituras de pseudopalavras

As métricas de avaliação coletadas para o item de pseudopalavras são a quantidade de palavras lidas (QPL) e a quantidade de palavras lidas corretamente (QPC) em um segmento de áudio de 60 segundos. Para cada pseudopalavra presente no teste, é informado se a mesma foi lida (correta ou incorretamente) ou não lida (sem tentativa de leitura da pseudopalavra). Uma leitura é classificada como correta se respeitar as regras que relacionam grafemas e fonemas na língua portuguesa. Para que uma leitura seja aprovada no item de pseudopalavras é preciso que mais que cinco pseudopalavras sejam lidas corretamente, isto é, a criança possui capacidade de decodificação mínima. Dependendo do objetivo da avaliação, outros testes podem ser utilizados para compor a avaliação de fluência leitora, como o uso de textos narrativos e questões de compreensão do texto.

O sistema para avaliação de leituras de pseudopalavras foi desenvolvido de forma a automatizar a coleta das métricas e informações citadas, reduzindo os custos associados à avaliação manual do item de pseudopalavras dentro de um processo em larga escala. Encontrar se uma palavra foi lida corretamente ou incorretamente é um dos maiores desafios para esse tipo de automatização. Como o teste é aplicado em crianças em fase de alfabetização, é comum a presença de leituras pausadas, falsos começos e repetições. Leituras que divergem da ordem esperada de leitura também são frequentes. Esses fatores dificultam que o sistema encontre ocorrências de determinadas pseudopalavras, tendendo a se distanciar do comportamento do ser humano.

3.2. Uso de modelos wav2vec2 para ASR

O *wav2vec2* é um modelo auto-supervisionado pré-treinado desenvolvido para aprender representações de fala. No aprendizado auto-supervisionado do *wav2vec2*, uma rede neural profunda é pré-treinada em dados não rotulados, onde representações contextuais da sequência de áudio são aprendidas. Essas representações pré-treinadas podem ser utilizadas em diversas tarefas para as quais não há dados suficientes disponíveis [Baevski et al. 2020, Vaessen and van Leeuwen 2021]. O processo de adaptação do modelo para uma determinada tarefa é chamado de *fine-tuning* e requer uma pequena quantidade de dados rotulados para o treinamento. [Baevski et al. 2020], por exemplo, realizou uma tarefa de reconhecimento que atingiu o estado da arte utilizando apenas 10 minutos de dados rotulados para realizar o *fine-tuning* do modelo.

Para que o ASR funcione adequadamente com o *wav2vec2*, a saída da rede pré-treinada com aprendizado auto-supervisionado é mapeada para uma outra camada, responsável por atuar como um classificador para as n letras do idioma que se deseja reconhecer. Para a língua portuguesa, cerca de trinta letras costumam ser utilizadas. Cada nó de saída da rede representa a probabilidade do modelo de uma letra ter ocorrido em certo instante do áudio (*frame*). Com isso, é possível associar a representação de fala aprendida pela rede neural e as probabilidades para cada nó de saída da rede com as letras do idioma, possibilitando o reconhecimento automático das palavras pronunciadas.

4. Descrição da solução para identificação de pseudopalavras em português

Para o reconhecimento automático de pseudopalavras foi utilizado um modelo pré-treinado[1], já adaptado para ASR pelo processo de *fine-tuning* em áudios da língua portuguesa utilizando o *corpus* CORAA [Junior et al. 2021].

[1] `https://huggingface.co/Edresson/wav2vec2-large-xlsr-coraa-portuguese`

4.1. Reconhecimento automático de pseudopalavras

O reconhecimento de pseudopalavras foi feito de duas formas: com e sem o auxílio de um modelo de língua. Modelos de língua são distribuições de probabilidades para sequências de símbolos (palavras ou letras). Na prática, estes modelos retornam a probabilidade de ocorrência de uma certa sequência de símbolos. No reconhecimento sem modelo de língua, o sistema se baseia somente nas características das representações de fala extraídas do áudio, ou seja, o sistema se apoia somente no modelo acústico para suas predições. Já no reconhecimento com o modelo de língua, o sistema faz uma ponderação entre os modelos acústico e de língua para predizer as pseudopalavras presentes em um áudio.

O modelo de língua utilizado neste trabalho atribui um peso para as pseudopalavras que se deseja reconhecer. Para isso, foi atributo o valor máximo de pesos aos unigramas e bigramas formados a partir da lista de pseudopalavras. Essa estratégia foi utilizada para forçar o reconhecimento a encontrar as pseudopalavras no áudio. Embora o modelo de língua force o reconhecimento para encontrar as pseudopalavras, tem-se que o mesmo não é determinante do conteúdo da transcrição com *wav2vec2*. A transcrição não conterá apenas as pseudopalavras e pode ocorrer a presença de outros termos caso o modelo acústico dê mais peso a esses termos durante o processo de reconhecimento.

4.2. Métodos para o cálculo do QPL

Para calcular a quantidade de palavras lidas (QPL), é preciso contabilizar os termos da transcrição que correspondem às pseudopalavras desejadas. Para isso, foram desenvolvidos dois métodos. O primeiro método, M^1_{QPL}, utiliza duas estratégias para alinhar a transcrição retornada pelo *wav2vec2* com a lista de pseudopalavras de referência.

Primeiramente, é utilizado um método para encontrar padrões de texto por aproximação. O algoritmo utiliza uma janela de contexto para aproximar o final da transcrição com algum ponto da lista de pseudopalavras. Um limiar de distância de edição é utilizado para determinar se as aproximações retornadas pelo algoritmo serão consideradas válidas. Quando o alinhamento encontra uma correspondência entre a transcrição e a lista de referência, um índice para a correspondência na lista de referência é retornado. O número de pseudopalavras na lista até o índice de correspondência é utilizado como o valor calculado para o QPL. Caso o alinhamento com essa estratégia falhe, um algoritmo de alinhamento entre o texto de hipótese e o texto de referência é empregado. O algoritmo retorna o número de correspondências, o número de inserções, o número de deleções e o número de substituições que foram necessários realizar para que o texto da hipótese (transcrição do *wav2vec2*) corresponda ao texto de referência (lista de pseudopalavras). Uma heurística que contabiliza o número de correspondências, substituições e deleções é utilizada para calcular o valor de QPL. A heurística soma o número de correspondências e de substituições com o número de deleções que formam grupos de até três deleções. As deleções são inseridas na contagem como uma forma de aproximar o número de pseudopalavras que podem ter sido puladas durante a leitura da criança.

O segundo método, M^2_{QPL}, calcula o QPL como sendo o índice da última pseudopalavra da lista de referência encontrada na transcrição. Para cada termo na transcrição, é verificado se o mesmo está presente na lista de referência. Caso esteja presente, o índice do termo na lista é retornado e comparado com o índice armazenado anteriormente. O índice de maior valor é retornado ao final da iteração e utilizado como valor de QPL.

4.3. Métodos para o cálculo do QPC

Para calcular a quantidade de palavras lidas corretamente (QPC), um algoritmo de alinhamento forçado de áudio foi utilizado para encontrar o conjunto de *frames* correspondentes a cada pseudopalavra ao longo da leitura. A probabilidade retornada pelo *wav2vec2* em predizer cada letra é uma informação que pode ser utilizada nesse processo. Baseado nos limites retornados pelo alinhamento forçado e na probabilidade atribuída na predição de cada letra é possível definir estratégias para avaliar a qualidade da pronúncia de uma pseudopalavra. A Equação 1 apresenta a fórmula utilizada neste trabalho para calcular a nota S da pronúncia de uma determinada palavra baseada em um conjunto de k letras l, que formam a palavra que se deseja reconhecer, e de k *frames* t, que representam o intervalo de tempo pelo qual as respectivas letras foram pronunciadas. Para cada pseudopalavra alinhada pelo algoritmo, essa fórmula é calculada e uma nota atribuída à pronúncia correspondente àquele trecho no áudio. Um limiar de 0.125 foi definido empiricamente para decidir se uma pronúncia está correta. Se o valor de S for menor que esse limiar, a leitura da pseudopalavra é considerada incorreta.

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

Duas estratégias foram utilizados para o cálculo do QPC. A primeira, M^1_{QPC}, utiliza a informação do QPL para cortar a lista de pseudopalavras na última palavra lida. A nova lista é passada para o algoritmo de alinhamento forçado e o valor S é calculado para cada pseudopalavra alinhada. O QPC do áudio é calculado como o número de pseudopalavras com valor S maior ou igual ao limiar. A segunda estratégia, M^2_{QPC}, utiliza toda transcrição para o processo de alinhamento forçado. Todos os termos presentes na transcrição são alinhados, porém somente os termos reconhecidos como pertencentes à lista de pseudopalavras são levados em consideração para o cálculo do QPC. Nesta estratégia, espera-se que o alinhamento forçado obtenha melhor resultado quando é recebida uma transcrição mais fiel ao que está sendo dito no áudio.

5. Resultados

Foram utilizados nos experimentos 1560 áudios de leituras de pseudopalavras do estado do Espírito Santo referentes a uma avaliação de fluência em larga escala aplicada em alunos do 2º ano do ensino fundamental. O conjunto de áudios contém amostras de cinco testes distintos de pseudopalavras contendo sessenta pseudopalavras cada. Três experimentos foram realizados para verificar a qualidade das estratégias para QPC e QPL discutidas anteriormente, além do uso do modelo de língua para o processo de reconhecimento. A Tabela 1 apresenta a definição de cada experimento realizado.

O experimento E_1 utilizou a transcrição sem o auxílio do modelo de língua. Os métodos M^1_{QPL} e M^1_{QPC} foram aplicados para calcular as métricas QPL e QPC, respectivamente. Nessa configuração, a transcrição sem modelo de língua é passada para o método M^1_{QPL}, que tenta encontrar as correspondências entre a transcrição e a lista de referência. O valor encontrado para o QPL é utilizado como entrada para o método M^1_{QPC}, que calcula o QPC a partir da contagem de pseudopalavras da lista cortada no

Tabela 1. Experimentos realizados

Experimento	Transcrito	Método QPC	Método QPL
E_1	**sem** modelo de língua	M^1_{QPC}	M^1_{QPL}
E_2	**com** modelo de língua	M^1_{QPC}	M^2_{QPL}
E_3	**com** modelo de língua	M^2_{QPC}	M^1_{QPL}

valor de QPL que estão acima do limiar de 0.125. Já os experimentos E_2 e E_3 utilizam a transcrição com o auxílio do modelo de língua. O experimento E_2 utiliza o mesmo método do experimento E_1 para o cálculo do QPC, se diferenciando pelo uso do método M^2_{QPL} para calcular o QPL como o índice da última pseudopalavra da lista encontrada na transcrição com o modelo de língua. Nos experimentos E_1 e E_2, o cálculo do QPC depende do valor de QPL encontrado pelo respectivo método utilizado para calcular a métrica. No experimento E_3, o cálculo do QPC independe do cálculo do QPL. Nesse experimento, a transcrição com modelo de língua é passada sem cortes para o algoritmo de alinhamento. As pseudopalavras da transcrição que pertencem à lista de referência são contabilizadas enquanto os termos da transcrição que não fazem parte da referência são ignorados (método M^2_{QPC}). O cálculo do QPL para o experimento E_3 utiliza o mesmo método do experimento E_1, mas utilizando a transcrição com o modelo de língua.

5.1. Análise de erro QPC e erro QPL

Quanto menor o erro dos métodos ao contabilizar as métricas QPL e QPC, melhor será a classificação automática sobre a qualidade da leitura. Nos experimentos E_1 e E_2, por exemplo, o cálculo do QPC depende do valor de QPL passado para o método M^1_{QPC}. Ainda, a classificação da leitura depende do número de pseudopalavras lidas corretamente. Portanto, erros no cálculo das métricas QPL e QPC podem impactar negativamente no desempenho automático. A Tabela 2 mostra o erro absoluto e o erro normalizado para as métricas QPC e QPL para cada um dos experimentos realizados. O erro absoluto é calculado como sendo a média do valor absoluto do erro, que é dado pela diferença entre o valor de QPC ou QPL predito automaticamente e o valor de referência do corretor humano. O erro normalizado representa o RMSE (raiz do erro quadrático medio), o qual permite analisar a distribuição dos erros, ou seja, se estão mais concentrados ou mais afastados do valor ideal (zero). Para ambas as medidas, quanto mais próximo de zero, melhor o desempenho da abordagem automática para contabilizar o QPC ou o QPL.

Tabela 2. Erro QPC e erro QPL

Métrica	E_1	E_2	E_3
Erro absoluto QPC	6,39	6,24	**2,75**
Erro absoluto QPL	4,64	8,56	**3,04**
Erro normalizado QPC	13,11	16,33	**6,56**
Erro normalizado QPL	3,90	5,58	**3,55**

O experimento E_3 obteve os menores valores de erro tanto para o cálculo do QPC quanto para o cálculo do QPL. Esse resultado indica que utilizar o modelo de língua

durante o reconhecimento auxilia positivamente na identificação das pseudopalavras lidas. O método M^1_{QPL} obteve um melhor resultado ao receber a transcrição com modelo de língua, encontrando mais facilmente as correspondências entre a transcrição e a lista de referência. Observando os erros QPL dos experimentos, é possível perceber que o método M^1_{QPL} foi superior ao método M^2_{QPL}, uma vez que o experimento E_2 exibiu maiores valores de erro absoluto e normalizado do que os erros dos demais experimentos. Para o cálculo do QPC, o método M^2_{QPC}, que utiliza a transcrição com modelo de língua sem cortes no algoritmo de alinhamento forçado, obteve o melhor resultado. Passar a transcrição inteira com o modelo de língua auxilia tanto na identificação das pseudopalavras presentes na lista de referência quanto melhora a qualidade do alinhamento, já que a transcrição passada é mais próxima do que foi dito no áudio.

O experimento que mostrou os melhores resultados foi o experimento E_3, que utiliza a transcrição com modelo de língua, o método M^1_{QPL} para calcular o QPL e o método M^2_{QPC} para calcular o QPC. Adicionar o modelo de língua durante o reconhecimento dos áudios trouxe melhores resultados para abordagem automática.

5.2. Análise da concordância com o corretor humano

Como discutido na Seção 3.1, para que uma leitura seja aprovada no teste de pseudopalavras é preciso que mais do que cinco pseudopalavras sejam lidas corretamente. Quando isso ocorre, é dito que uma leitura pertence à classe positiva. Caso contrário, é dito que a leitura pertence à classe negativa. A Tabela 3 mostra o desempenho da abordagem automática na classificação das leituras para os experimentos realizados.

Tabela 3. Acurácia, precisão e revocação na classificação das leituras

Métrica	E_1	E_2	E_3
Acurácia	0,86	0,88	**0,91**
Precisão	0,93	0,87	**0,97**
Revocação	0,90	**0,99**	0,92

A precisão indica a proporção de todas as predições atribuídas à classe positiva que pertencem a essa classe de acordo com o corretor humano. Quanto maior o valor da precisão, maior é a concordância da abordagem automática com o corretor humano ao dizer que uma leitura foi aprovada no teste de pseudopalavras. A revocação indica a proporção das leituras classificadas como positiva pelo corretor humano que são preditas como leituras pertencentes a essa classe. Quanto maior o valor da revocação, mais leituras acima do limiar são classificadas corretamente pela abordagem automática. A acurácia pode ser entendida como uma medida geral da concordância das predições automáticas com o corretor humano, mostrando a proporção de predições em acordo com a avaliação humana. O experimento E_3 obteve o melhor resultado de precisão e de acurácia, enquanto o experimento E_2 obteve o melhor resultado para a revocação. Com o experimento E_3 foi obtida uma precisão de $0,97$ e uma acurácia de $0,91$. Esses valores mostram que essa configuração obteve a maior concordância com o corretor humano, sendo o experimento E_3 o que produziu os resultados mais consistentes.

6. Conclusões

Este trabalho teve como contribuição: (i) uma nova abordagem para a avaliação automática de leituras de pseudopalavras utilizando treinamento com áudios não rotulados e mostrando que é possível alcançar bons resultados na avaliação automática utilizando o modelo para reconhecer leituras de pseudopalavras de crianças em fase de alfabetização; (ii) o desenvolvimento de dois métodos para o cálculo da métrica QPL, dois métodos para o cálculo da métrica QPC e um modelo de língua para auxiliar no reconhecimento das pseudopalavras presentes no teste; (iii) diferente de estudos passados, foi utilizado um conjunto considerável de áudios reais para validação da proposta.

O desempenho da abordagem automática para cada configuração de experimento foi aferido por meio de uma comparação com os dados anotados pelo corretor humano utilizando leituras de crianças do 2º ano do ensino fundamental do Espírito Santo. O experimento E_3 obteve os melhores resultados, tanto em termos do erro no cálculo do QPC e do QPL quanto na concordância com o corretor humano em classificar as leituras de pseudopalavras. Analisando os dados, foi possível concluir que a adição de um modelo de língua durante o reconhecimento dos áudios trouxe resultados melhores na identificação das pseudopalavras lidas. Os métodos M^1_{QPL} e M^2_{QPC} obtiveram os melhores resultados para o cálculo do QPL e do QPC.

Os resultados encontrados demonstram que a solução pode ser utilizada em testes em larga escala e apoiar aplicações para avaliação de qualidade de leitura de pseudopalavras em jogos educacionais, como em [Ceccon and Porto 2020, Hautala et al. 2020, Silva et al. 2019]. A proposta apresentou uma precisão de 0,97 e uma revocação de 0,92, portanto percebe-se que a solução um gera um bom indicador de boas leituras (verdadeiros positivos), embora possa gerar ligeiramente mais erros na classe de negativos (inclusão de falsos negativos). Apesar dos bons resultados, mais esforços precisam ser feitos para que o erro no cálculo das métricas QPC e QPL sejam reduzidos. Como discutido na Seção 5.1, melhorias no cálculo do QPC e do QPL melhoram diretamente a qualidade da classificação das leituras, uma vez que essa classificação depende das métricas citadas. Diversos fatores influenciam nestes erros, como a presença de palavras não esperadas, erros na aplicação dos testes, presença de ruídos ou outras vozes ao longo da gravação. Técnicas para filtrar ruídos ou para identificação de falas anômalas ao teste podem aumentar a qualidade da solução. Ainda, como a leitura é classificada positiva ou negativa de acordo com o QPC maior que 5, áudios com valores próximas desse limiar estão mais sujeitos à classificação incorreta. Para adoção da tecnologia, a seleção desses áudios para homologação pelo corretor manual é uma boa prática e aumenta a qualidade do resultado. Esse conjunto geralmente equivale de 3% a 7% dos áudios das aplicações realizadas pela Fundação CAEd.

Por fim, embora a solução tenha sido projetada para o português brasileiro, é possível sua utilização para outras nações de língua portuguesa utilizando conjuntos de treinamento para essas variações do idioma. É necessário, contudo, maior investigação para verificar a qualidade da solução em outras línguas.

Referências

Almeida Silva, W., Carchedi, L., Gomes Jr, J., Souza, J., Barrere, E., and Souza, J. (2021). A framework for large-scale automatic fluency assessment. *International Journal of*

Distance Education Technologies, 19.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *10.48550/ARXIV.2006.11477*. arXiv.

Batista, A. A. G. (2011). Alfabetização, leitura e ensino de português: desafios e perspectivas curriculares. *Revista Contemporânea de Educação*, 6(12):246–272.

Ceccon, D. L. and Porto, J. B. (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. SBC.

Cresswell, J., Schwantner, U., and Waters, C. (2015). *A Review of International Large-Scale Assessments in Education*. PISA, The World Bank, Washington, D.C./OECD Publishing, Paris.

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.

Dias, N. M., León, C. B. R., Pazeto, T. d. C. B., Martins, G. L. L., Pereira, A. P. P., and Seabra, A. G. (2016). Avaliação da leitura no brasil: Revisão da literatura no recorte 2009-2013. *Psicologia: teoria e prática*, 18(1):113–128.

Duchateau, J., Cleuren, L., Van hamme, H., and Ghesquière, P. (2007). Automatic assessment of children's reading level. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1:1210–1213.

Elenius, D. and Blomberg, M. (2004). Comparing speech recognition for adults and children. *Proceedings of FONETIK 2004*, pages 156–159.

Farrall, M. L. and Ashby, J. (2019). The role of assessment in structured literacy. *Perspectives on Language and Literacy*, 45(3):31–35.

Hautala, J., Heikkilä, R., Nieminen, L., Rantanen, V., Latvala, J.-M., and Richardson, U. (2020). Identification of reading difficulties by a digital game-based assessment technology. *Journal of Educational Computing Research*, 58(5):1003–1028.

Junior, A. C., Casanova, E., Soares, A., de Oliveira, F. S., Oliveira, L., Junior, R. C. F., da Silva, D. P. P., Fayet, F. G., Carlotto, B. B., Gris, L. R. S., and Aluísio, S. M. (2021). Coraa: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese.

Mechelli, A., Gorno-Tempini, M. L., and Price, C. J. (2003). Neuroimaging studies of word and pseudoword reading: consistencies, inconsistencies, and limitations. *Journal of cognitive neuroscience*, 15(2):260–271.

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National ....

Pantoja, J., Sousa, A., and de Araújo Júnior, R. M. (2018). Alfa autista: uma aplicação mobile para o auxílio na alfabetização do autista através de método fônico. um es-

tudo de caso na apae-marabá. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1873.

Passos, C., Fernandes, I., and Goldschmidt, R. (2019). Elaboração e avaliação de projeto de aprendizagem apoiado em jogos educacionais digitais: Um relato de experiência com alunos em alfabetização. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 674.

Pinheiro, Â. M. V. and de Araújo Vilhena, D. (2022). Teste de reconhecimento de palavras e pseudopalavras: validades de conteúdo e externa. *Signo*, 47(88):145–161.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., and Perdigão, F. (2017a). Automatic Evaluation of Children Reading Aloud on Sentences and Pseudowords. In *Proc. Interspeech 2017*, pages 2749–2753.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., and Perdigão, F. (2017b). Detection of Mispronunciations and Disfluencies in Children Reading Aloud. In *Proc. Interspeech 2017*, pages 1437–1441.

Proença, J. D. L. (2018). *Automatic assessment of reading ability of children*. PhD thesis, Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Rasinski, T. V. (2004). Assessing reading fluency. *Pacific Resources for Education and Learning (PREL)*.

Salles, J. F. d. and Parente, M. A. d. M. P. (2007). Avaliação da leitura e escrita de palavras em crianças de 2ª série: abordagem neuropsicológica cognitiva. *Psicologia: Reflexão e Crítica*, 20(2):220–228.

Shivakumar, P. G. and Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer speech & language*, 63:101077.

Silva, W. A., Gomes Jr, J., Knop, I., Barrére, E., and Souza, J. (2019). Talk2me: Uma abordagem computacional para auxiliar na identificação de falhas no processo de alfabetização. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 723.

Vaessen, N. and van Leeuwen, D. A. (2021). Fine-tuning wav2vec2 for speaker recognition. *arXiv preprint arXiv:2109.15053*.

Wagemaker, H. (2014). International large-scale assessments: From research to policy. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, pages 11–36.

Yılmaz, E., Pelemans, J., and Van hamme, H. (2014). Automatic assessment of children's reading with the flavor decoding using a phone confusion model. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.