Avalia Online: um sistema para avaliação em larga escala de testes de fluência de leitura

Luiz Carlos Carchedi, Eduardo Barrére, Jairo Francisco de Souza

[1]LApIC Research Group, Programa de Pós-Graduação em Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF) - MG - Brasil

{lcarchedi,eduardo.barrere,jairo.souza}@ice.ufjf.br

Abstract. *Reading assessment requires specific methods, which are costly because, in most cases, they involve specialist human evaluators. To reduce the cost of large-scale assessments, it is necessary to automate parts of the process. In this paper, the Evaluate Online System is presented, which enables the automation of process steps and the reduction of the time needed to obtain results. To guarantee a low error rate, the system allows the selection of the subset with a high probability of error to undergo human verification. Our approach achieved a classification accuracy of 95.97% on a real dataset containing more than 9,000 readings.*

Resumo. *A avaliação da leitura requer métodos específicos, os quais são custosos por, na maioria das vezes, envolver avaliadores humanos especialistas. Para redução do custo em avaliações em larga escala, faz-se necessário automatizar partes do processo. Nesse artigo, é apresentado o Sistema Avalia Online, que possibilita a automatização de etapas do processo e a redução do tempo necessário pra obtenção de resultados. Para garantia de baixa taxa de erros, o sistema permite a seleção do subconjunto com alta probabilidade de erro para passar por verificação humana. O sistema alcançou 95,97% de acurácia na tarefa de classificação em uma base de dados real com mais de 9 mil leituras.*

1. Introdução

A comunicação é essencial no desenvolvimento do aluno em seus primeiros anos de escola. Dentre seus fatores, a linguagem oral (oralidade) possui destaque por permitir ao aluno a construção de laços com o professor e com os demais alunos, possibilitando seu desenvolvimento social e a produção de conhecimento [Chaer and Guimarães 2012]. A linguagem oral também age como um facilitador na aprendizagem da escrita e do letramento infantil eficazes e significativos para a criança [Alves 2019]. Um dos aspectos que compõem a oralidade é a fluência na fala e na leitura [Biemiller 1977].

A avaliação dos alunos em diferentes competências possibilita o levantamento de dados para a obtenção do panorama geral do seu desenvolvimento, permitindo a aplicação de políticas públicas voltadas para a melhoria do ensino. Essas avaliações precisam ser aplicadas em larga escala e padronizadas, aliadas às tecnologias que permitam a avaliação automatizada, fornecendo assim informações úteis aos professores de forma contínua, consistente e acessível [Black et al. 2011]. Avaliações dessa natureza geralmente são realizadas por meio de provas escritas, impossibilitando a avaliação de algumas competências que são consequentemente negligenciadas, como é o caso da avaliação da leitura.

DOI: 10.5753/sbie.2021.217139

A automatização da avaliação da fluência na leitura pressupõe a criação de um sistema que possa avaliar a leitura humana presente em um áudio gravado, utilizando reconhecimento automático de fala (do inglês *Automatic Speech Recognition* - ASR): um conjunto de técnicas com o objetivo de gerar texto a partir de um sinal de áudio [Gruhn et al. 2011] e que possibilitam a automatização do processo de avaliação da leitura. Essa automatização traz benefícios, como a redução do custo com a avaliação (menor necessidade de mão de obra), e a redução significativa no tempo necessário para geração dos resultados da avaliação. Para a automatização da avaliação da oralidade, as medidas de fluência de um minuto podem ser utilizadas como classificadoras e são tecnicamente importantes por demandar pouco espaço de armazenamento e pouco custo de transmissão dos áudios, o que faz com que a métrica seja comumente utilizada por diversos autores [Deeney 2010, Valencia et al. 2010]. Contudo, a geração dessas métricas não é trivial, uma vez que necessitam de ferramentas especializadas, capazes de lidar com falantes de baixa faixa etária [Yeung and Alwan 2018] que, por não terem o trato vocálico completamente desenvolvido quando muito jovens, apresentam características de voz específicas, e com as interferências presentes em áudios provenientes de gravações em escolas por conta de uso de equipamentos de baixa qualidade, sons ambientes, etc.

O presente trabalho propõe a automatização de parte significativa da avaliação da oralidade infantil, entretanto, dadas as limitações e dificuldades na automatização do processo, é apresentada uma estratégia para alcançar alta acurácia na avaliação, separando os resultados entre os que apresentam confiabilidade aceitável e os que precisam ser avaliados manualmente. Como resultado dessa estratégia, é disponibilizada uma ferramenta para apoiar avaliações de fluência em larga escala que não dispensa completamente o uso da avaliação manual, mas faz com que esta seja significativamente reduzida, melhorando o tempo e o custo da avaliação de todo o conjunto de áudios.

O trabalho está organizado da seguinte forma. A Seção 2 contempla os trabalhos relacionados que apresentam soluções computacionais para análise de fluência em leitura. A Seção 3 aborda o processo de automatização da avaliação de fluência utilizando a metodologia do CAEd e como um *framework* foi definido para esse contexto. Em seguida, na Seção 4, o sistema implementado foi avaliado com um conjunto expressivo de áudios coletados em escolas brasileiras de crianças e os resultados mostram a baixa taxa de erro da solução e discute o ganho de tempo da solução em comparação à análise com avaliadores humanos. Por fim, as conclusões e trabalhos futuros são discutidos na Seção 5.

2. Trabalhos Relacionados

O uso de técnicas de ASR na educação tem trazido resultados importantes na área. Em [Xie et al. 2012] é destacado o uso de ASR para a avaliação automatizada de proficiência em segunda língua, enquanto [Neumeyer et al. 1996] apresentam um sistema que se utiliza de técnicas de ASR para atribuir notas às pronúncias dos estudantes. Os resultados desse trabalho corroboram a hipótese de que estudantes mais fluentes tendem a falar mais rápido que os iniciantes. De forma similar, [Cucchiarini et al. 2000] discutem resultados das avaliações de fluência realizadas por especialistas e os compara àqueles obtidos automaticamente, concluindo que o número de palavras por período de tempo é a métrica que mais refletiu o resultado da avaliação realizada por especialistas.

Recentemente houve avanços também na avaliação de fluência em língua portu-

guesa. Em [Carchedi et al. 2018a] é proposto o uso de ASR para a automatização do processo de avaliação da fluência em leitura, porém ainda com uma taxa de erro muito alta. Em [Gomes Jr et al. 2019], prova-se que a utilização de algoritmos de alinhamento forçado em conjunto com um sistema de ASR levam a uma maior taxa de acurácia quando a máquina tem conhecimento do texto de referência. Estes trabalhos apresentam resultados importantes, mas ainda não permitem a sua utilização na prática, pois percebe-se uma taxa de erro alta nas instâncias com desempenho próximo ao limite das classes de fluentes e não fluentes (ponto de corte das classes).

Uma outra vertente de trabalho, apresentada em [Silva et al. 2019], discute o uso de sistemas ASR para gamificação da avaliação de leituras em português do Brasil. Através de um jogo de celular, o aluno é incentivado a ler e a gravar sua leitura. Os resultados são avaliados automaticamente e um professor pode receber os resultados para interpretação do desempenho de cada aluno. Da mesma forma que em trabalhos anteriores, a avaliação é totalmente automática, fazendo com que a interpretação dos resultados tenha que ser feita considerando possíveis erros da máquina.

Para alcançar uma alta confiabilidade de avaliações em larga escala, este trabalho apresenta o sistema Avalia Online, que implementa a avaliação de leitura em língua portuguesa utilizando métricas de fluência. Diferente dos trabalhos da literatura, a alta confiabilidade do sistema é alcançada através de uma avaliação automática dos áudios e de uma avaliação manual do subconjunto de instâncias que se encontram perto de uma região passível de uma taxa de erro elevada. Assim, a dupla checagem desses áudios garante resultados mais confiáveis para os gestores. Além disso, o sistema automatiza também diversas tarefas do processo de avaliação, desde a coleta até a entrega dos resultados através de um *dashboard* que permite um uso mais prático para usuários interessados na avaliação que, muitas vezes, não possuem conhecimento técnico de TI. Ao longo do artigo, é apresentada a arquitetura dessa solução para larga escala e é discutido como a avaliação manual, embora mais custosa, leva a um custo-benefício aceitável, uma vez que é realizada em um subconjunto muito pequeno da base original.

3. Processo de automatização de avaliação em larga escala de leitura oral

Este projeto foi desenvolvido para atender as avaliações realizadas pelo Centro de Políticas Públicas e Avaliação da Educação (CAEd/UFJF). O CAEd é uma empresa pública que mensura o rendimento de estudantes em 15 estados do Brasil para auxiliar ações de melhoria da qualidade da educação. Para avaliação de fluência na leitura, os alunos são avaliados a partir de um conjunto de textos e classificados como fluentes ou disfluentes, de acordo com as métricas obtidas de um conjunto de leituras com duração de um minuto. Entre as métricas utilizadas em cada item do teste, estão a quantidade de palavras lidas corretamente (QPC), o total de palavras lidas (QPL) e a precisão na leitura (relação QPC / QPL). O leitor fluente consegue ler, no mínimo, 65 palavras por minuto com 90% de precisão. Essas métricas eram extraídas manualmente por avaliadores humanos, acarretando alto custo e demora na entrega do resultado para análise dos dados e geração de relatórios.

A classificação da fluência baseada nessas métricas pode ser extraída automaticamente das leituras, o que motivou o desenvolvimento do sistema **Avalia Online** com o objetivo de auxiliar no processo de correção dos testes de leitura de textos. O objetivo

do sistema é permitir uma extração com alta acurácia das métricas necessárias e, quando necessário, apoiar avaliadores humanos na avaliação manual de áudios de baixa qualidade. Baseado em um sistema de reconhecimento automático de fala (ASR) e de um algoritmo que alinha os fonemas com o que deveria ser lido no texto de referência, o sistema verifica se as palavras foram ou não pronunciadas de maneira correta de acordo com sua transcrição fonética. As métricas para a classificação geram uma fronteira entre os possíveis resultados na avaliação do leitura, e, nas proximidades dessa fronteira, pode-se identificar uma região de incerteza do sistema, onde pequenas variações nas métricas obtidas podem afetar o resultado da avaliação do áudio. Essas variações podem ser resultado de características do áudio como, por exemplo, o aplicador da avaliação falando junto ao leitor, a qualidade do áudio gerar um volume baixo, a existência de barulho/ruídos ao fundo, chiados no áudio, pessoas falando ao fundo ou mesmo o som abafado.

Os áudios distantes dessa fronteira possuem uma confiabilidade alta no resultado apresentado pelo sistema, enquanto aqueles próximos a ela têm classificações que não são tão confiáveis e, por esse motivo, necessitam passar por uma avaliação manual, realizadas por avaliadores capacitados e preparados para esse tipo de tarefa. Assim, os resultados obtidos através de avaliações manuais servem tanto para a definição dos áudios que ficaram próximos à fronteira entre os possíveis resultados, como também servem para testar a confiabilidade do sistema.

Considerando as possibilidades de erros na automatização do processo, é importante levar em consideração a precisão (porcentagem de elementos classificados corretamente) dos resultados apresentados pelo sistema. Tendo uma alta precisão, a confiabilidade do sistema permitirá que as leituras que de fato pertencem ao conjunto de fluentes anteriormente definido não precisem passar por avaliações manuais. As demais leituras deverão ser encaminhadas à avaliação manual.

A principal vantagem trazida por uma ferramenta para a automatização de um processo é a rapidez na coleta dos dados no uso de grandes amostras, o menor custo de administração e taxas de retorno mais expressivas [Dixon 2001]. Assim, através da ferramenta os avaliadores podem encontrar de maneira mais fácil e rápida os pontos a serem corrigidos e os gestores do ensino podem voltar sua atenção a esses pontos.

Na utilização do sistema, o professor coleta através de um aparelho celular, os áudios com as leituras realizadas pelos alunos e os envia, através de um *app*, para o sistema de avaliação, o qual extrai as características de cada áudio e gera as métricas. Utilizando um extrato avaliado manualmente, o administrador ajusta os parâmetros a serem utilizados na avaliação e verifica a acurácia atual do sistema, realizando simulações para encontrar os melhores parâmetros a serem utilizados na classificação e os aplica no restante da base. Os áudios são, então, classificados como *Fluente* ou como *Não-fluente*. Caso identifique áudios muito próximos de uma região de baixa acurácia, o administrador pode enviar uma amostra desses áudios à avaliação manual para garantir a alta qualidade dos dados gerados através do processo de dupla checagem. Caso existam dúvidas durante a avaliação manual, os avaliadores podem destacá-las para que sejam auxiliados por outros avaliadores antes de um parecer final sobre as métricas daquele áudio, permitindo uma avaliação cooperativa e um consenso em casos inesperados. A qualquer momento, o administrador do sistema pode se utilizar do *dashboard* para que obtenha as estatísticas gerais do sistema, bem como as estatísticas separadas por projetos, por textos, por avali-

adores e etc, possibilitando a tomada as decisões e ajustes da melhor configuração para classificação dos áudios. A Figura 1 representa a arquitetura do sistema, dividida em módulos independentes que separam suas funcionalidades.

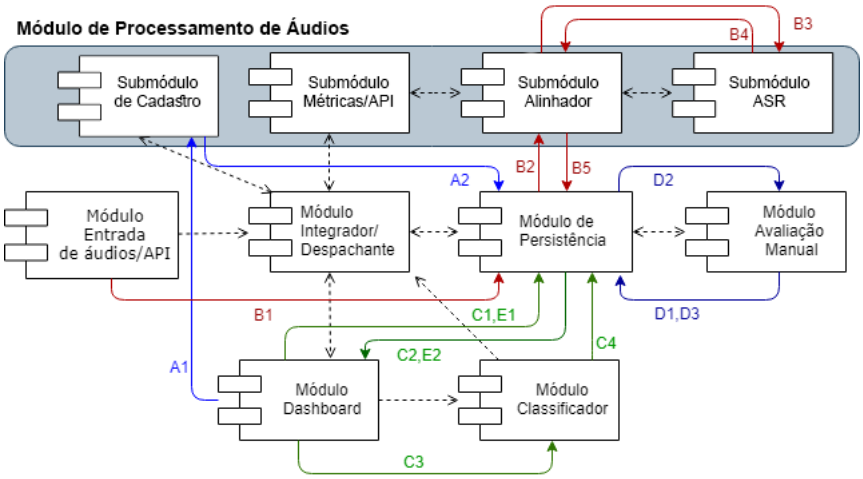

Figura 1. Arquitetura do Sistema Avalia Online

O Módulo *Dashboard* cadastra um projeto [A1] no Sistema e as informações pertinentes do projeto são enviadas ao Módulo de Persistência [A2]. Em seguida, os áudios são inseridos no sistema através do Módulo de Entrada de Áudios/API, e associados a um projeto existente [B1]. No Módulo de Processamento de áudios, é realizada a a transcrição do áudio e, ao final, são retornadas ao Módulo de Persistência as métricas obtidas.

O Módulo de Processamento de Áudios é um sistema auxiliar, composto pelos submódulos: Cadastro, Métricas, Alinhador e ASR. Esse sistema é responsável pela transcrição do sinal de áudio gerando as métricas de cada arquivo.

O submódulo de Cadastro é onde são cadastrados as diretrizes (nomeadas como 'Projetos') utilizadas. Nessas diretrizes são especificadas as regras, texto de referência, dicionário e etc. a serem utilizados na transcrição de um áudio. No submódulo ASR é onde, utilizando as diretrizes do projeto, são realizadas as transcrições dos áudios. O submódulo alinhador, realiza o alinhamento forçado da transcrição realizada com o texto referente àquele áudio. Finalmente, no submódulo Métricas, a partir da comparação entre a transcrição realizada e o texto de referência, é possível a obtenção das métricas como: quantidade de palavras lidas por minuto, palavras lidas corretamente e precisão.

Na etapa seguinte, a partir da consulta ao Módulo de persistência [C1], o Módulo *Dashboard* obtém os dados a respeito das transcrições e a partir destes, exibe simulações sobre os áudios de um determinado projeto. Essas simulações permitem encontrar os melhores parâmetros a serem usados pelo Módulo Classificador, que define quais áudios serão enviados à avaliação manual. Dando continuidade, o Módulo Avaliação Manual requisita ao Módulo de Persistência [D1] um áudio e o recebe em [D2]. Realizada essa avaliação, as métricas manuais são retornadas ao Módulo de Persistência [D3]. Ao final do processo, o Módulo *Dashboard* faz uma requisição ao Módulo de Persistência [E1] e recebe [E2] os dados que são exibidos em forma de gráficos, tabelas e outras interfaces que fornecem uma visão generalizada do projeto e permitem tomadas de decisões.

Para facilitar o compartilhamento de dados com outras aplicações educacionais, o sistema utiliza a ontologia Onto4LA [Carchedi et al. 2018b], a qual é baseada no modelo de eventos, facilitando análises em aplicações de *learning analytics*. A Onto4LA é uma ontologia que usa um modelo de proveniência para dados educacionais, facilitando a integração dos dados de diferentes sistemas computacionais e, com isso, auxiliando na análise dos dados. No Avalia Online, os dados são gerados utilizando o vocabulário da Onto4LA e pode ser consumido por sistemas que interpretem esses dados.

O Módulo *Dashboard* é voltado para os supervisores que, através dele, podem criar novos projetos, obter uma visão geral do sistema (assim como de cada um dos projetos cadastrados) e, além disso, realizar simulações. Tais simulações são realizadas para a classificação das leituras dentro de um projeto, apresentando dados de acurácia e gráficos de dispersão, gráficos da curva ROC e matriz de confusão com os resultados. O Módulo *Dashboard* permite ainda a visualização do histórico de uso dos avaliadores, a quantidade de avaliações realizadas de maneira geral e individual, a média de avaliações nas últimas semanas e participação na parte colaborativa. A Matriz de Confusão, assim como os gráficos de erro, considera apenas os áudios que possuem métricas obtidas manualmente e faz uma relação entre as classificações segundo as métricas geradas automaticamente pelo sistema e segundo as métricas resultantes de avaliações manuais. A Matriz de Confusão exibe quantos áudios, segundo os parâmetros informados, se encontram dentro de cada possibilidade de classificação. A partir desses instrumentos, o supervisor tem condições de encontrar as melhores métricas para a classificação daquela base e pode aplicá-las aos demais áudios. Todos os módulos do sistema se comunicam através do Módulo Integrador, permitindo que haja a troca de informações necessária.

A arquitetura do sistema, intencionalmente desenvolvida como módulos independentes, tem por objetivo a separação das funcionalidades do sistema, o que torna mais fácil sua configuração e manutenção. Essa independência entre os módulos também permite que hajam diferentes configurações do sistema a partir da união de mais módulos, por exemplo: permitindo um poder de processamento maior com a classificação automática realizada por mais módulos Classificadores ao mesmo tempo, ou ainda tornando possível que sejam desenvolvidas e testadas diferentes ferramentas para a avaliação manual sem que isso comprometa o restante do sistema. A partir dessa arquitetura também se torna viável a inserção de outros módulos que dariam novas funcionalidades ao sistema, como por exemplo módulos para avaliar leituras de pseudo-palavras (itens comumente utilizados para testar competência de leitura no método fônico) e para cruzar resultados de diferentes testes em uma mesma classificação, módulos para integração dos dados das avaliações com outras ferramentas avaliativas, entre outros.

4. Experimentos

Foi realizado um experimento com dados reais de avaliações de fluência de leitura para aferir a qualidade da avaliação automática de acordo com as avaliações manuais (confiáveis). As avaliações manuais foram realizadas por um grupo externo de professores do ensino fundamental contratados e treinados pelo CAEd/UFJF. Os professores informaram no sistema a quantidade de palavras lidas pelo aluno e os erros de leitura. Uma amostra dessas avaliações passaram por rechecagem para garantir a qualidade dos dados. Esse experimento utiliza uma base que possui 9412 áudios contendo a leitura de crianças nos dois primeiros anos de alfabetização, coletados em escolas no Espírito Santo,

Pará, Paraíba e Pernambuco, a qual chamaremos de DFLU. A base foi fornecida pelo CAEd/UFJF e, por confidencialidade dos dados, não foi fornecida nenhuma informação que pudesse caracterizar individualmente cada áudio, como localidade, idade, gênero dos alunos ou dos avaliadores. Embora essa limitação impeça análises mais detalhadas dos dados, a solução projetada leva em consideração apenas a voz humana, não sendo necessário informar características do indivíduo para o classificador.

As medidas de qualidade definidas para os experimentos são: a diferença entre as métricas encontradas (erros de QPL e QPC) e a acurácia da classificação dos áudios nas classes *Fluente* e *Não Fluente*. Vale ressaltar que as medidas de erro e de acurácia trazem informações diferentes da qualidade do sistema, visto que, ainda que haja um valor alto de erro nas métricas, esse erro não necessariamente vai influenciar na classificação binária que o sistema realiza quando o erro se dá em leituras muito boas ou muito ruins. A classificação binária é um requisito do CAEd/UFJF para a classificação da leitura de textos narrativos, refletindo um modelo de avaliação amplamente utilizado na literatura da avaliação de fluência (ARTIGOS). O *Dashboard* oferece as ferramentas para que o supervisor possa calibrar as métricas que melhor classificam os áudios e devem ser utilizadas para a classificação de toda a base, ajustando o melhor ponto de corte para a classificação e análise de curva ROC.

Todos os áudios foram submetidos a um modelo acústico treinado com aproximadamente 80 horas de áudios de leituras infantis que não fazem parte desse experimento. Foram inseridos 15 minutos de áudios de leituras muito ruins para melhor generalizar o modelo. Foram utilizados 80% para treinamento e 20% para validação para treinar um modelo HMM-GMM de trifones com SAT (*speaker adaptive training)*, o qual foi utilizado para treinar uma *time-delay neural network* (TDNN) contendo 7 camadas com 384 neurônios, uma camada de saída com 512 neurônios, e uma função objetivo LF-MMI (*Lattice-free Maximum Mutual Information)*. A rede foi treinada por 25 épocas e obteve uma taxa de erro de palavras de 14,95.

No início do experimento, 5% dos áudios foram escolhidos aleatoriamente e submetidos à avaliação de especialistas para extração das métricas. Esses áudios formam uma base de referência que contém as avaliações automáticas assim como as avaliações dos especialistas. A Figura 2 mostra os gráficos gerados pela ferramenta, os quais representam a distribuição de erro entre as avaliações automáticas e as manuais. O erro de um áudio é a diferença na contagem das palavras obtidas automaticamente e manualmente(confiável), e esses erros são calculados para o QPL e para o QPC. Através desses gráficos é possível notar que a maior parte dos áudios apresenta erro igual a 0, o que significa que na maioria das vezes o sistema conseguiu extrair a métrica de maneira correta.

Nas avaliações realizadas pelo CAEd/UFJF, as métricas utilizadas para a classificação de um áudio como Fluente são *QPC = 65* e *Precisão = 90%*. A base DFLU classificada com esses parâmetros, apresentou um resultado de 2915 áudios (30,97%) fluentes e 6497 áudios (69,03%) não-fluentes. Em comparação com a base de referência, houve uma acurácia de 95,54%. Para definir o melhor ajuste para a classificação, o que é comum quando a amostra dos dados pode ter características levemente distintas, o *Dashboard* apresenta uma curva ROC para auxiliar o supervisor. A figura 3 apresenta a curva ROC gerada para a base DFLU. Seguindo a metodologia apresentada em [Carchedi et al. 2018a], foi escolhido o ponto da curva com a maior taxa de *verdadeiros*

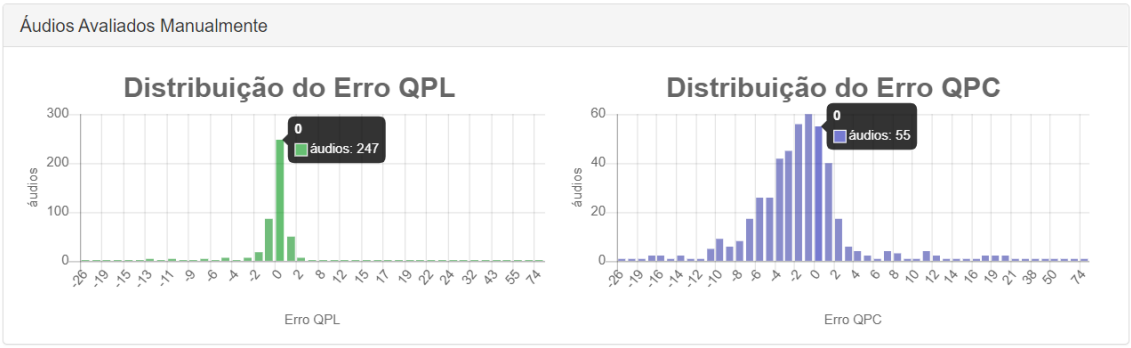

Figura 2. Distribuição de erro entre as avaliações automáticas e as manuais

positivos e a menor taxa de *falsos positivos*, o que aumenta a confiabilidade do sistema em relação àqueles áudios classificados como fluentes. Dessa maneira, a curva ROC gerada para a base DFLU com 5% de seus áudios avaliados manualmente aponta para os melhores parâmetros como sendo uma Precisão de 90% e um QPC de 66 para a melhor classificação. Ao utilizar esses valores, a acurácia do sistema subiu para 95,97%.

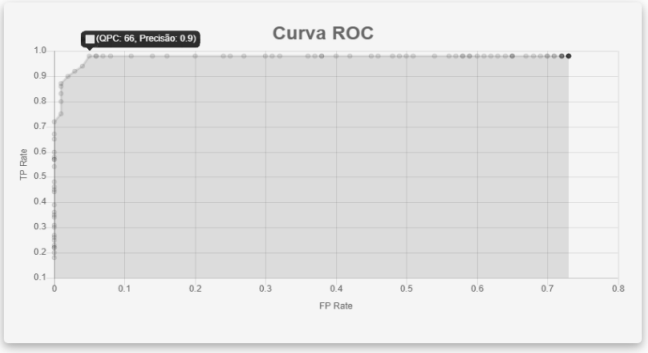

Figura 3. Simulação da base DFLU com todos os áudios avaliados manualmente

Os experimentos demonstraram que o sistema é capaz de alcançar alta acurácia. O tempo médio de avaliação de cada áudio foi de 67,43 segundos em uma máquina de poder de processamento mediano, enquanto uma avaliação humana demora, em média, 228 segundos[1]. O sistema permite a configuração de novas instâncias do Módulo de Processamento dos Áudios de forma a distribuir os áudios em diversos nós de processamento, conforme a infra-estrutura disponível. Com esse valor médio de processamento, pode-se comparar o ganho de tempo em uma avaliação em larga escala. Segundo os dados disponíveis no portal do INEP[2] a respeito da Provinha Brasil, no ano de 2018 a prova foi realizada por 5.201.730 alunos. Considerando que esses alunos realizassem a avaliação de fluência, na qual cada aluno grava 3 áudios de um minuto na metodologia aplicada pelo CAEd/UFJF, seriam gerados 15.605.190 áudios para serem avaliados, resultando em aproximadamente 260 mil horas de gravações que necessitariam ser avaliadas manualmente. Considerando-se que, para garantir a qualidade do sistema para diferentes formas

[1]O tempo médio da avaliação humana foi calculado à partir de dados de avaliações anteriores e é calculado desde o início da reprodução do áudio até a submissão dos dados pelo avaliador.

[2]http://provabrasil.inep.gov.br/microdados

de pronúncia (sotaques), seria necessário classificar manualmente apenas 5% dos áudios (em torno de 13 mil horas), reduzindo consideravelmente o esforço humano.

5. Conclusões

O presente trabalho apresentou uma solução computacional para apoiar avaliações de fluência em larga escala. Entre os diferenciais do trabalho, estão (1) o uso, em um mesmo ambiente, de avaliação automática por aprendizagem de máquina e de avaliação manual por especialistas da área de forma a aumentar a confiabilidade dos dados gerados pela solução; (2) uma arquitetura flexível para comportar avaliações de um volume alto de avaliações diárias, o que permite a sua aplicação em um ambiente real; e (3) uma avaliação da acurácia da solução automática utilizando dados reais de avaliações de fluências ocorridas em quatro estados da Federação.

O esforço manual é significativamente reduzido, melhorando assim o tempo e o custo da avaliação de todo o conjunto de áudios. Vale ressaltar que o uso de avaliadores manuais não é um requisito do sistema, mas uma garantia de homologação dos dados gerados automaticamente e uma facilidade para geração de mais dados para treinamento do classificador. Levando-se em consideração as dimensões do país, com suas variações fonéticas, a dificuldade de controlar o espaço de aplicação de provas e o nível de ruído desses espaços, além da alta confiabilidade dos resultados que esse tipo de solução necessita gerar, torna-se imprescindível o uso de dados manuais para melhorias de classificação a cada nova avaliação, como em diferentes anos letivos e diferentes regiões do país.

A solução permite a reprodutibilidade dos experimentos com diferentes parâmetros, além de uma base de dados padronizada que pode ser utilizada para comparação e geração de análises entre avaliações e para treinamento de outros modelos. O aspecto modular da solução permite que esta seja adaptada para outros tipos de modelos e avaliações. Desta forma é possível adaptar novas instâncias de módulos, com novos comportamentos, de acordo com a infraestrutura disponível, como módulos para paralelizar o processamento dos áudios, ou um Dashboard para uma ferramenta voltada pra outro objetivo específico da área de ciência de dados.

Os resultados apresentados, contudo, apresentam limitações que podem ser exploradas em novos estudos. A falta de características dos alunos e dos avaliadores não permite aprofundar em análises que possam extrair mais conhecimento dos dados. A falta de uma base de dados pública para avaliar fluência é um dos problemas da área e são necessários esforços nesta direção. Os trabalhos existentes utilizam bases proprietárias e, como é o caso deste trabalho, geralmente apenas um subconjunto dos dados é fornecido. Ainda, a classificação binária visa reproduzir a metodologia do CAEd/UFJF, baseada na literatura, e comparar com a avaliação humana, mas novas classes poderiam ser criadas no *framework* para reconhecer novos grupos de leitores.

Como trabalhos futuros, estuda-se a distribuição do sistema com licença BSD para que outros centros possam fazer uso da solução e façam adaptações para seus cenários de aplicação. Outros modelos acústicos serão treinados e disponibilizados na solução para permitir a avaliação de pronúncia de sílabas, o que pode permitir resultados mais acurados em outros tipos de avaliação da competência leitora, como em Testes de Competência de Leitura de Palavras e Pseudopalavras (TCLPP).

Referências

Alves, S. S. S. (2019). O papel da oralidade para o desenvolvimento do letramento em crianças da educação infantil. Seminário Interlinhas, 7(1):211–215.

Biemiller, A. (1977). Relationships between oral reading rates for letters, words, and simple text in the development of reading achievement. Reading Research Quarterly.

Black, M. P., Kazemzadeh, A., Tepperman, J., and Narayanan, S. S. (2011). Automatically assessing the abcs: Verification of children's spoken letter-names and letter-sounds. ACM Transactions on Speech and Language Processing (TSLP), 7(4):15.

Carchedi, L. C., Soares, E., Gomes Jr, J., Barrére, E., and Souza, J. (2018a). Avaliação automática da fluência em leitura para crianças em fase de alfabetização. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), volume 29, page 11.

Carchedi, L. C., Souza, J., Barrére, E., and Mendonça, F. (2018b). Onto4la: uma ontologia para integração de dados educacionais. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação, volume 7, page 439.

Chaer, M. R. and Guimarães, E. d. G. A. (2012). A importância da oralidade: educação infantil e séries iniciais do ensino fundamental. Disponível em: `http://pergaminho.unipam.edu.br/documents/43440/43870/a-importancia.pdf`. Acesso em: 04 abril 2018.

Cucchiarini, C., Strik, H., and Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. The Journal of the Acoustical Society of America, 107(2):989–999.

Deeney, T. A. (2010). One-minute fluency measures: Mixed messages in assessment and instruction. The Reading Teacher, 63(6):440–450.

Dixon, J. (2001). Evaluation tools for flexible delivery (workshop version). Melbourne: TAFE frontiers.

Gomes Jr, J., Silva, W. A., Souza, J., Barrére, E., and Souza, J. (2019). Uso de alinhadores forçados para avaliação automática em larga escala da fluência em leitura. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), volume 30, page 61.

Gruhn, R. E., Minker, W., and Nakamura, S. (2011). Statistical pronunciation modeling for non-native speech processing. Springer Science & Business Media.

Neumeyer, L., Franco, H., Weintraub, M., and Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. In International Conference on Spoken Language. ICSLP 96., volume 3, pages 1457–1460. IEEE.

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.

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., and Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. Reading Research Quarterly, 45(3):270–291.

Xie, S., Evanini, K., and Zechner, K. (2012). Exploring content features for automated speech scoring. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 103–111. Association for Computational Linguistics.

Yeung, G. and Alwan, A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. In Proc. Interspeech 2018, pages 1661–1665.