UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Luiz Carlos Carchedi

SISTEMA COLABORATIVO PARA AVALIAÇÃO DE TESTES DE FLUÊNCIA EM LARGA ESCALA

Juiz de Fora

2019

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Luiz Carlos Carchedi

SISTEMA COLABORATIVO PARA AVALIAÇÃO DE TESTES DE FLUÊNCIA EM LARGA ESCALA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Eduardo Barrére

Coorientador: Jairo Francisco de Souza

Juiz de Fora

2019

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Carchedi, Luiz Carlos.
Sistema Colaborativo para Avaliação de Testes de Fluência em Larga Escala / Luiz Carlos Carchedi. – 2019.
62 f. : il.

Orientador: Eduardo Barrére
Coorientador: Jairo Francisco de Souza
Dissertação (Mestrado Acadêmico) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2019.

1. Análise de aprendizagem. 2. Automação da avaliação. 3. Avaliação em larga escala. 4. Reconhecimento automático de fala I. Barrére, Eduardo, orient. II. Souza, Jairo Francisco de, coorient. III.Título.

Luiz Carlos Carchedi

SISTEMA COLABORATIVO PARA AVALIAÇÃO DE TESTES DE FLUÊNCIA EM LARGA ESCALA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 31 de Julho de 2019

BANCA EXAMINADORA

Prof. Dr. Eduardo Barrére - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Jairo Francisco de Souza - Coorientador
Universidade Federal de Juiz de Fora

Prof. Dra. Liamara Scortegagna
Universidade Federal de Juiz de Fora

Prof. Dr. Diego Dermeval Medeiros da Cunha Matos
Universidade Federal do Alagoas

Trabalho dedicado à minha mãe,
que tornou tudo isso possível.

AGRADECIMENTOS

Ao final de mais uma etapa, faz-se necessário o reconhecimento daqueles que tornaram possível a realização da mesma.

Certamente a liberdade está entre as cosias mais valiosas que existem, e a liberdade só é alcançada através das oportunidades. Graças as oportunidades que recebi, hoje essa etapa pode ser concluída.

Em primeiro lugar, agradeço à minha mãe, que com uma força admirável e uma retidão exemplar, fez com que eu tivesse as oportunidades que ela sonhou pra si; é graças a ela que tudo isso acontece e, sem qualquer dúvida, é dela essa vitória.

Em seguida, um agradecimento muito especial à Daniele por toda atenção, carinho, compreensão, apoio, incentivo e dedicação, estando comigo todo o tempo, desde a sugestão de fazer o curso, ao treinamento da última apresentação. Através dela, pude ver o mundo de uma perspectiva melhor. Certamente cada vitória tem mais sentido quando temos com quem compartilhá-la.

Agradeço àqueles da minha família que me apoiaram e sempre me ajudaram em tudo que precisei.

Agradeço ainda ao professor Marcelo Moreno, cujo apoio e direcionamento foi fundamental pra que eu ingressasse nesse curso.

Para sempre serei grato aos professores Eduardo Barrére e Jairo Souza, que mesmo já havendo alcançado a excelência em suas carreiras, ainda assim, com paciência e perseverança me acompanharam todo o tempo. Agradeço a cada correção recebida e por continuamente me ajudarem, desde o início da graduação até aqui, propiciando minha formação e me dando diversas oportunidades para o meu crescimento.

O fato de haver professores como esses, mostra que ainda existem motivos para nos esforçarmos e sermos melhores. É lamentável que o nosso país não valorize os nossos mestres da maneira que seria justa.

Ainda em tempo, expresso minha gratidão aos colegas do LApIC, que incontáveis vezes me ajudaram e sempre o fizeram de bom grado, em especial, ao Warley e ao João Vitor, que me ajudaram diretamente e muito na realização desse trabalho, tornando-o viável.

Por fim, agradeço à toda a equipe do PGCC pelos serviços prestados.

"Em algum lugar, algo incrível está
esperando para ser descoberto."
Carl Sagan

RESUMO

A avaliação da aprendizagem é de suma importância no processo de ensino e pode ser realizada de diferentes maneiras a depender de seu objetivo. Dentre as avaliações pertinentes no início da vida escolar, a linguagem oral se destaca, uma vez que é fundamental para o desenvolvimento do aluno e sua interpretação do que escuta, organização de pensamentos e mesmo para a formulação de suas ideias.

Existem também as avaliações voltadas para o levantamento de dados e informações para a obtenção do panorama geral do desenvolvimento dos alunos em uma área, região ou até mesmo todo o país. Essas avaliações precisam ser aplicadas em larga escala e seus resultados permitem a aplicação de políticas públicas voltadas para a melhoria do ensino.

As avaliações aplicadas em larga escala, devido às suas características, em geral são realizadas por meio de provas escritas, o que impossibilita a avaliação de algumas competências, que consequentemente são negligenciadas nessas avaliações, como é o caso da linguagem oral.

O presente trabalho apresenta o sistema **Avalia Online**, desenvolvido com o objetivo de possibilitar a avaliação em larga escala da oralidade de crianças em fase de alfabetização a partir da leitura de textos, utilizando a avaliação automática. Como a avaliação da oralidade de maneira automática possui limitações, o sistema faz uso de uma abordagem colaborativa permitindo que avaliadores capacitados avaliem as leituras onde o sistema apresenta incertezas e que contribuam entre si nas dúvidas que eventualmente ocorram. O sistema foi avaliado a partir de dois experimentos: o primeiro voltado á avaliação do classificador automático e o segundo voltado á avaliação da abordagem colaborativa. A partir desses experimentos, foram obtidos resultados que comprovam a importância do sistema proposto.

Palavras-chave: Análise de aprendizagem. Automatização da avaliação. Avaliação em larga escala. Reconhecimento automático de fala.

ABSTRACT

The proper assessment of learning is of paramount importance to an efficient teaching process. Assessments can be done in many different forms depending on what aspects of the learning process are under evaluation. Oral language stands out at the beginning of school life since students at that age have not yet fully developed their written skills, making oral expression fundamental for the development of students as it is the means through which they will receive information, interpret that information, organize their thoughts based on that information and even formulate new ideas.

Besides the evaluation of each student individually, there are tests applied on a large scale which aim at data-gathering for obtaining the overall picture of the education system in an given area of knowledge or geographic region in the local or the national level. The results of these assessments allow the decision makers to develop and implement public policies aimed at improving the school system. On the other hand, given the peculiarities of oral language, large-scale tests are usually carried out in written form, completely neglecting competences which are exclusive of oral communication.

With that in mind, the present paper introduces the **Avalia Online**, an automated system designed to evaluate, in large scale, the orality of children in the literacy phase. As it is with any automatic system, Avalia Online has its limitations, nonetheless it uses a collaborative approach that allows trained evaluators to intervene when needed, making the assessment process more reliable as a whole.

Key-words: Learning analysis. Assessment automation. Large scale assessment. Automatic speech recognition.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ASR	Reconhecimento Automático de Fala
CAEd	Centro de Políticas Públicas e Avaliação da Educação
ENEM	Exame Nacional do Ensino Médio
FOAF	Friend of a Friend
IPA	International Phonetic Alphabet
JSON	Javascript Object Notation
LApIC	Laboratório de Aplicações e Inovação em Computação da UFJF
LDB	Diretrizes e bases da educação nacional
PAEBES	Programa de Avaliação da Educação Básica do Espírito Santo
PLN	Processamento de Linguagem Natural
PNC	Parâmetros Nacionais Curriculares
PNE	Plano Nacional de Educação
QPC	Quantidade de Palavras Lidas Corretamente
QPL	Quantidade de Palavras Lidas
ROC	Receiver Operation Characteristic
SAEB	Sistema de Avaliação da Educação Básica
SAETHE	Sistema de Avaliação Educacional de Teresina
SEM	Simple Event Model
SGBD	Sistema de Gestão de Banco de Dados
SKOS	Simple Knowledge Organization System
UFJF	Universidade Federal de Juiz de Fora

SUMÁRIO

1 INTRODUÇÃO

Em todo o processo de ensino, a avaliação da aprendizagem é fundamental, uma vez que por meio dela é possível acompanhar o desenvolvimento dos alunos. As avaliações quando ocorrem em larga escala são essenciais para a definição de estratégias e políticas públicas que visem o aumento da qualidade do ensino em um âmbito geral [45].

A partir de avaliações em larga escala, os educadores podem obter informações suficientes sobre turmas, escolas, regiões e até mesmo todo o país, possibilitando a definição de metas e planos de ensino para o aperfeiçoamento do aprendizado [46], tal como o planejamento de um modelo de ensino mais eficaz que permita melhorar os indicadores de interesse relacionados à educação [2].

Contudo, a avaliação da aprendizagem é uma questão complexa e existem diferentes posicionamentos baseados em vertentes distintas, fazendo com que hajam posturas variadas a respeito de como a aprendizagem deve ser avaliada, dificultando a chegada a um consenso [22]. Em [48], a avaliação da aprendizagem é apresentada sob duas vertentes: a primeira delas se baseia nas Diretrizes e Referencial Curricular para a Educação Infantil junto com as Diretrizes e Parâmetros do Ensino Fundamental; a segunda, por sua vez, se baseia nas Diretrizes e Parâmetros do Ensino Médio e Estudos em Avaliação Educacional, nos processos avaliativos existentes no Sistema Educacional Brasileiro: Saeb e Enem.

Ainda assim, existe uma base consensual comum onde a ideia de qualidade na educação se fundamenta. Essa base leva em consideração dois pontos: (1) as Leis de Diretrizes e Bases da Educação Nacional (LDB)[21] e (2) o Plano Nacional de Educação (PNE) [32][19].

Nas LDB são estabelecidos parâmetros como carga horária e condições de trabalho dos professores, carga horária dos alunos dentro da escola, a maneira com a qual o ensino será ministrado, o que será assegurado pelo sistema de ensino aos educandos com necessidades especiais e ainda a valorização dos profissionais da educação de acordo com suas especializações, piso salarial aos mesmos e período reservado ao estudo, planejamento e avaliação inclusos na carga horária de trabalho [32]. No Plano Nacional de Educação, por sua vez, são estabelecidos os objetivos e metas para a educação infantil, educação fundamental e educação do ensino médio. No PNE são estabelecidos padrões mínimos de infra-estrutura para o funcionamento adequado das instituições de educação públicas e privadas, que respeitando as diversidades regionais, assegurem o atendimento das características das distintas faixas etárias e das necessidades do processo educativo.

Dentre as competências ensinadas aos alunos nos seus primeiros anos de alfabetização, está a linguagem oral, que possui destaque, uma vez que toda a produção do conhecimento parte desse aspecto da comunicação [15]. A importância da linguagem oral se relaciona com o desenvolvimento da criança em seus primeiros anos de escola, dessa

maneira, ela é uma ferramenta fundamental para que a criança construa laços com o professor, com os demais alunos e também se desenvolva também socialmente. Portanto, torna-se imprescindível que o professor atue de maneira intencional para o desenvolvimento dessa linguagem [20].

Quando bem desenvolvida, a oralidade (linguagem oral) permitirá que o aluno interprete o que ouve, tenha pensamento organizado e responda perguntas com lógica e clareza. Assim, é de fundamental importância que se dedique um tempo especial nas tarefas diárias escolares para o desenvolvimento da oralidade [15]. Um dos aspectos que compõem a oralidade é a *fluência* na fala e na leitura [4]. Isto posto, é crucial que haja um tempo especial nas tarefas escolares para o desenvolvimento dessas habilidades e sua avaliação [30]. O desenvolvimento da fluência na leitura permite que a criança tenha uma maior capacidade interpretativa do conteúdo do texto, uma vez que é necessário menor esforço para a decodificação do texto escrito, esse esforço é direcionado para a interpretação do texto [42].

Para a decisão entre a mais eficaz entre as diferentes formas de avaliar o aprendizado do aluno, é necessário levar em consideração quais os itens a serem avaliados e qual a escala de avaliação. Uma avaliação tradicional geralmente é aplicada por um professor em uma sala de aula, mas quando se trata de uma avaliação aplicada em larga escala, existem órgãos especializados em sua realização e outros meios se fazem necessários.

A avaliação em larga escala é um dos principais instrumentos para a elaboração de políticas públicas dos sistemas de ensino e redirecionamento das metas das unidades escolares. Seu foco é o desempenho da escola e o seu resultado é uma medida de proficiência. Essa medida possibilita aos gestores a implementação de políticas públicas eficazes, e possibilita às unidades escolares um retrato de seu desempenho [8].

No Brasil, para obter informações sobre o nível de alfabetização das crianças que estão começando o segundo ano da vida escolar e conseguir essas informações em larga escala, o Governo Federal implementou a Provinha Brasil[1] no ano de 2008. Ao realizar essa avaliação, a intenção é que os dados aferidos sirvam como uma amostragem da habilidade leitora do aluno e que, ao analisá-los, o professor tenha orientação para aperfeiçoar sua forma de ensino [29].

As avaliações em larga escala possuem características específicas, demandando cuidados na preparação desse tipo de avaliação. Por esse motivo, tais avaliações têm um caráter completamente técnico, possibilitando atender a uma grande quantidade de alunos e fazendo com que seja necessária uma correção padronizada. Desse modo, a avaliação perde a propriedade de mostrar quais as deficiências no aprendizado de um aluno específico, e passa a mostrar o resultado de forma estatística [50].

[1] http://portal.inep.gov.br/web/guest/provinha-brasil

Avaliações em larga escala, de maneira geral, ocorrem por meio de provas escritas que são aplicadas nacionalmente por agências de avaliação cujo o objetivo é mensurar a competência dos alunos, tais avaliações são baseadas no que é estabelecido pelo Governo Federal através do Ministério da Educação. Entretanto, nem todas as competências podem ser avaliadas através de provas escritas, como no caso da oralidade, onde o foco principal se encontra na maneira em que o aluno utiliza a linguagem oral para se comunicar; isso faz com que essa competência seja sensivelmente mais difícil de ser avaliada em larga escala pois torna-se um processo custoso em termos financeiros e de tempo. Dessa maneira, devido a impossibilidade de avaliação da oralidade através do meio escrito e as dificuldades de implementação em larga escala desse tipo de avaliação, as provas de avaliações da Educação Básica, nacionais e internacionais, como o Sistema de Avaliação da Educação Básica (SAEB) e a Provinha Brasil, em sua grande maioria, são puramente escritas [18]; isso tem como consequência um desequilíbrio que faz com que a avaliação da linguagem oral não receba toda a atenção e preparo devidos, fazendo muitas vezes que nem mesmo seja considerada [18], portanto tornam-se interessantes as propostas de soluções mais baratas e mais rápidas para esse processo, o que traria como consequência direta a possibilidade da sua aplicação em escalas maiores.

Atualmente, existem diferentes estudos e experiências voltadas à avaliação da educação infantil devido à sua importância e possibilidades e também por haverem pontos em aberto. Em [5] os autores dão importância à disfluências como hesitações, sussurros, latências alongadas, entonações de perguntas entre outras, e também desenvolvem um sistema para a automatização da avaliação de crianças em leituras de palavras em voz alta; entretanto, nesse trabalho, o foco é na leitura de palavras isoladas, mas não na leitura de textos, enquanto o presente trabalho tem seu foco na avaliação da leitura de textos.

Em [26] é apresentado o contexto das correções *automatizadas* de avaliações de leituras orais por parte de leitores iniciantes. O trabalho leva em conta dados a respeito do desempenho dos leitores e também considera o efeito potencial dos erros que acontecem durante a leitura. Nesse trabalho, porém, não são apresentados os meios pelos quais são abordados os erros provenientes das correções *automatizadas*.

Para que a avaliação da linguagem oral também seja levada em consideração, é necessário que hajam meios pelos quais ela também seja realizada em larga escala. Considerando-se as características específicas da avaliação da oralidade, faz-se necessário a preparação de avaliações próprias para essa modalidade. No Brasil já existem empresas desenvolvendo um procedimento voltado especificamente para a realização da avaliação da oralidade, como por exemplo, o Centro de Políticas Públicas e Avaliação da Educação (CAEd) da Universidade Federal de Juiz de Fora. No caso da avaliação de algumas competências como a fluência em leitura, para serem realizadas em larga escala, é necessário o uso de tecnologias de apoio, e uma das maneiras pelas quais isso pode ser obtido é

através de soluções computacionais, que também são eficazes em relação ao volume de dados gerados por tais avaliações quando aplicadas em larga escala.

Uma vez adotadas soluções computacionais para a resolução dos problemas da escalabilidade, um segundo problema a ser considerado é a confiabilidade do sistema no processo de automatização da avaliação da oralidade. No reconhecimento de fala podem ocorrer erros devido a uma série de características do sistema que fazem com que sua confiabilidade seja variável, como por exemplo: a qualidade do áudio, ruídos durante a gravação, o ambiente no qual o áudio foi captado etc. A confiabilidade do sistema é importante porque erros na avaliação têm impacto direto nos resultados e consequentemente, na tomada de decisões a respeito de políticas e estratégias para o desenvolvimento da oralidade em crianças em fase de alfabetização, dessa maneira, geralmente tais avaliações são realizadas manualmente para garantir a corretude de seus dados.

Existem métodos para a minimização de problemas causados por esses erros, e através desses métodos são buscadas soluções nas quais os áudios captados sejam mais próximos a realidade por possuírem menos interferências. Dessa maneira é desejável o fomento de soluções que levem em consideração políticas eficazes que consigam um melhor resultado no desenvolvimento da oralidade infantil.

Além dos problemas citados anteriormente que podem afetar na avaliação das leituras, as variações da maneira de falar causadas pelo sotaque de uma região também causam impacto nos resultados obtidos com o reconhecimento de fala e isso pode afetar sensivelmente o resultado esperado, esse é outro problema a ser considerado, em especial no Brasil devido à sua grande área territorial, que faz com que em diferentes regiões hajam diferentes sotaques e características regionais, como consequência, ao avaliar o áudio relacionado a uma leitura, o sistema também precisa levar esse aspecto em consideração.

1.1 OBJETIVOS E PRODUTO

O objetivo desta dissertação é fornecer uma abordagem computacional de baixo custo operacional, que garanta uma alta acurácia da avaliação da fluência em leitura, viabilizando sua aplicação em larga escala. A solução leva em conta o Reconhecimento Automático de Fala (ASR) para a extração de métricas a partir das gravações de leituras realizadas por crianças em fase de alfabetização. A abordagem aplica técnicas de ASR para classificação automática de áudios como fluentes ou não fluentes.

Como objetivos específicos, este trabalho fornece uma solução que, por um lado, minimiza a intervenção de avaliadores humanos no processo avaliativo e, por outro lado, oferece ferramentas que auxiliam no processo de avaliação humana.

Como produto do trabalho, foi desenvolvido um sistema de apoio para configuração de avaliações em larga escala e controle do processo de avaliação automática e humana,

chamado **Avalia Online**. A abordagem foi avaliada com uma base real contendo áudios de avaliações realizadas pelo CAEd em quatro estados brasileiros. Os resultados mostram que a abordagem atinge os requisitos e pode ser utilizada em avaliações futuras.

1.2 ORGANIZAÇÃO DA DISSERTAÇÃO

Para demostrar os resultados alcançados por esse trabalho, a dissertação está organizada da seguinte maneira: o capítulo 2 discute os conceitos relacionados à avaliação da fluência em larga escala; no capítulo 3 é apresentado o sistema desenvolvido para a avaliação em larga escala. No capítulo 4 são descritos os experimentos realizados e, finalmente, no capítulo 5, são apresentadas as considerações finais e previsões de trabalhos futuros.

2 AVALIAÇÃO DA FLUÊNCIA EM LARGA ESCALA

Neste capítulo são apresentados os conceitos relacionados à avaliação da fluência em larga escala. Na seção 2.1 é discutida a Avaliação da Linguagem Oral. A seção 2.2 apresenta o reconhecimento automático de fala, fundamental para a automatização da avaliação da linguagem oral. Finalmente, a seção 2.3 apresenta os desafios enfrentados para a automatização da avaliação.

A linguagem oral, fundamental na formação do aluno, tem suas particularidades que tornam necessária uma avaliação específica para a mesma [37]. De acordo com [15], entre os recursos utilizados para a avaliação da linguagem oral infantil estão as atividades apresentadas oralmente, as conversas informais entre alunos e professores e ainda as observações de conversas espontâneas entre os próprios alunos. O desenvolvimento dessa linguagem por parte do aluno em fase de alfabetização, acarretará na fluência na leitura [42], e dessa forma, a avaliação da leitura por parte do aluno em fase de alfabetização serve como parâmetro para a avaliação do desenvolvimento da linguagem oral do mesmo.

Apesar de diferentes definições para fluência na leitura, cada uma das quais colocando ênfase variável em seus componentes, parece haver um consenso crescente de que precisão, automaticidade e prosódia são seus componentes fundamentais [38]. Os resultados obtidos nas avaliações da leitura mostram uma relação direta entre a fluência e a compreensão do texto lido [42].

Em [44], é apresentado um trabalho que desenvolve um sistema automatizado para a avaliação da linguagem oral de crianças de escolas primárias que tem o inglês como a segunda língua. O trabalho leva em conta a prosódia e usa um sistema para a detecção de proeminência de palavras com base em características prosódicas. Nesse trabalho, procura-se utilizar-se métodos para a supressão de ruído e não se utiliza da avaliação manual em qualquer momento. No presente trabalho, a avaliação manual é utilizada quando ruídos são identificados nos áudios que contém a leitura, gerando uma classificação confiável mesmo para esses áudios.

2.1 AVALIAÇÃO DA LINGUAGEM ORAL

Esse trabalho leva em conta, para fins de orientação, os Parâmetros Nacionais Curriculares (PNC) [9] do Ministério da Educação. Esse documento define os parâmetros a serem utilizados na educação infantil a fim de que os estudantes atinjam os patamares de qualidade de aprendizado previamente estabelecidos e que sejam alcançadas as metas para a educação como um todo. Especificamente voltados para a fluência na leitura e escrita, existem os “Parâmetros Nacionais Curriculares da Língua Portuguesa”. Esses parâmetros utilizam uma estrutura que divide o ensino da língua portuguesa em ciclos, onde cada ciclo possui avaliações e objetivos específicos que procuram preparar o aluno para o ciclo

seguinte [9].

Para avaliações que levem em conta uma escala maior, se faz necessária a padronização da avaliação da oralidade. Essa padronização aliada a tecnologias que permitam a avaliação automatizada, fornece informações úteis aos professores de forma repetida, consistente e acessível [6].

Levando em consideração os PNC, é possível estipular uma forma padronizada para a avaliação da leitura, procurando estabelecer critérios que sejam bem definidos e através destes, obter resultados que direcionem o posicionamento do avaliador em relação ao aluno avaliado.

Nos PNC da língua portuguesa, o "Primeiro Ciclo" é voltado para os alunos no início de sua alfabetização, e ao terminá-lo, é esperado que o aluno tenha desenvolvido habilidades específicas: (1) narrar histórias conhecidas e relatos de acontecimentos, mantendo o encadeamento dos fatos e sua sequência cronológica, ainda que com ajuda do professor. Outra habilidade esperada é a de (2) demonstrar compreensão do sentido global de textos lidos em voz alta. Espera-se ainda que o aluno (3) consiga ler de forma independente textos cujo conteúdo e forma são familiares. Por fim, o aluno deve conseguir (4) escrever utilizando a escrita alfabética de acordo com a convenção ortográfica, demonstrando preocupação com a segmentação do texto em palavras e em frases [9].

Para a avaliação da linguagem oral baseada nos PNC, são utilizadas diferentes métricas. Em [47] são apresentadas algumas métricas como: palavras lidas por minuto, palavras lidas corretamente por minuto, precisão, prosódia e compreensão. Esse estudo também aponta uma relação entre a fluência e a compreensão do texto. Em [25] os autores utilizam essas métricas para classificar as crianças em grupos que expressam sua capacidade de leitura.

As medidas de fluência de um minuto podem ser utilizadas como classificadoras e são tecnicamente interessantes por demandar pouco espaço de armazenamento dos áudios e pouco custo de transmissão dos arquivos, o que faz com que a métrica seja comumente utilizada por diversos autores [23, 47]. Para esses autores, esta métrica é útil para identificar estudantes com dificuldades de leitura e definir a intervenção do professor ou uma nova política de ensino.

As ferramentas projetadas para a avaliação da leitura devem observar o desempenho dos leitores-alvo utilizando-se dos tipos de textos adequados [39]. Os resultados obtidos com essas ferramentas permitem que as métricas sejam geradas para a avaliação da leitura. Tais métricas possibilitam encontrar anomalias na leitura e avaliá-la como fluente/disfluente. Entre as anomalias mais comuns estão hesitações, sussurros, alongamentos, entonação de perguntas, entre outras [5].

2.2 AUTOMATIZAÇÃO DA AVALIAÇÃO

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. Nesse cenário, é utilizado um sistema de Processamento de Linguagem Natural (PLN). Um sistema PLN leva em conta a estrutura da linguagem para abranger aspectos da comunicação humana como a fala, as palavras, os textos e as sentenças considerando todo o contexto no qual está inserido [27]. O presente trabalho tem seu foco na avaliação da leitura de textos e as sentenças considerando todo o contexto no qual está inserido [27]. Entre as aplicações na área de PLN, está a técnica de *reconhecimento automático de fala*.

2.2.1 **Reconhecimento automático de fala**

O reconhecimento automático de fala (*Automatic Speech Recognition* - ASR) é um conjunto de técnicas que tem como objetivo gerar texto a partir da fala contida em um sinal de áudio [33]. Na abordagem mais utilizada, as palavras são inferidas através de uma análise do sinal de áudio e por meio do cálculo da probabilidade de uma dada palavra (ou sequência de palavras) estar associada a um trecho de áudio de acordo com o seu som e sua frequência de ocorrência naquele idioma, além de também ser levado em consideração as propriedades acústicas das unidades de fala (em geral fonemas) e do conhecimento do idioma para a consideração de quais palavras são seguidas por quais palavras [33]. Para que isso aconteça, é utilizada uma ferramenta chamada *Alinhador Forçado*. Um Alinhador Forçado tem por objetivo alinhar temporalmente o que é pronunciado com um texto de referência, de maneira que apenas as palavras de interesse sejam reconhecidas e transcritas nos momentos corretos.

Avanços significativos já foram alcançados no reconhecimento de fala para adultos e muitas vezes é assumido que esses avanços serão transferidos para o domínio com crianças, todavia essa transferência não é tão automática e, quando o sistema é treinado especificamente para o uso com crianças, tende a ter resultados significativamente melhores e, por esse motivo, é adequado que o sistema leve em consideração a faixa etária do falante [34, 53]. Em [52], foi demonstrado que o treinamento do modelo acústico especializado para a faixa etária que se deseja reconhecer a fala sempre obtém melhores resultados que treinamentos genéricos. Esse fato foi corroborado posteriormente em [43], onde os autores treinaram modelos acústicos com falas pertencentes a grupos de crianças de diferentes faixas de idade e obtiveram sempre melhores resultados quando o conjunto de testes pertencia ao mesmo grupo dos dados de treinamento. À medida que as investigações acerca de ASR em crianças foram avançando, também foram surgindo formas de lidar com os desafios envolvidos nessa tarefa.

Entre os trabalhos da literatura que investigam o uso de ASR para voz infantil é consenso que a acurácia do reconhecimento de fala infantil geralmente é mais baixa do

que em adultos [16], isso se deve às características específicas existentes na fala infantil, como por exemplo os efeitos causados pelas mudanças nas vozes das crianças a partir dos 6 anos de idade. Essas diferenças estão relacionadas, principalmente, às distinções anatômicas e morfológicas em relação ao trato vocálico, dessa maneira crianças têm maiores frequências fundamentais e variabilidade no espectro da voz [31]. Outra diferença comumente apresentada na literatura diz respeito às habilidades linguísticas infantis, principalmente em relação à dificuldade de pronúncia de certos fonemas. A dificuldade em fazer um modelo específico pra cada faixa etária se relaciona com a escassez de bases de treinamento, fazendo com que apenas um modelo genérico sem especificidade de faixa etária seja criado, mesmo assim a modelagem de pronúncia pode aumentar a acurácia desses reconhecedores de fala infantil quando o mesmo tem seu modelo acústico treinado com adultos [16], e a utilização de bases de treinamento composta por áudios contendo fala de pessoas adultas pode obter bons resultados no reconhecimento de fala infantil através da aplicação de algumas técnicas, como a normalização pelo tamanho do trato vocálico, regressão linear de máxima verossimilhança, treinamento adaptativo por falante, máxima probabilidade a posteriori, entre outras [36]. Também é preciso considerar as diferenças entre os áudios de leitura e de fala espontânea, para que o reconhecimento a ser realizado seja adequado, por isso, uma forma de melhorar a acurácia desses reconhecedores é através da adição das formas alternativas de pronúncia [34].

2.2.2 **Aplicações para o reconhecimento automático de fala**

Existem várias abordagens na literatura que utilizam o ASR como uma ferramenta para avaliação automática de leitura e fala. Muitas dessas abordagens utilizam sistemas de ASR para atribuir pontuação à fala de forma automatizada e, dessa forma, conseguir avaliar a proficiência de pessoas em um dado idioma [41, 17, 54].

O trabalho de [55] destaca que, uma vez que o ASR é cada vez mais usado nas mais diferentes aplicações, é fundamental que os sistemas que se utilizam de ASR estejam preparados para lidar com os diferentes sotaques encontrados dentro de um mesmo idioma; neste trabalho são apresentadas algumas técnicas utilizadas para melhorar o reconhecimento mesmo em casos que tenham sotaques carregados e proposta de uma abordagem que combina essas técnicas e consegue uma melhoria significativa no reconhecimento de fala. Em [1, 35] vemos que o sotaque é um dos fatores-chave na variabilidade a qual o reconhecimento de fala está sujeito. Em [35] os autores lidam com os sotaques utilizando-se de análises estatísticas, e quando uma certa quantidade de dados de adaptação estava disponível, a modelagem do dicionário de pronúncia foi adotada para reduzir os erros de reconhecimento causados por diferenças de pronúncia. Quando um grande corpus foi coletado para cada tipo de sotaque, os modelos dependentes de sotaque foram treinados e um sistema de identificação de sotaque foi desenvolvido para a seleção do modelo. O artigo relata resultados experimentais para os dois esquemas e sua eficiência em cada situação. Já

no trabalho [3] o reconhecimento de padrões é utilizado para a identificação dos sotaques brasileiros tal como as demais variações regionais da fala do português brasileiro, e no trabalho [51] é abordada a diferença de sotaque no inglês falado em diferentes partes do mundo. Nesse trabalho é proposto um método que, através do reconhecimento automático de fala, permite identificar qual o sotaque presente na fala contida no sinal de áudio. O trabalho traz a ideia de cada sotaque pertencer a uma região do planeta e, a partir de oito sotaques principais, os demais sotaques podem ser identificados. O trabalho apresenta uma acurácia de 80% na identificação dos sotaques, e mostra que, com o uso desse método, sistemas de reconhecimento de fala têm uma melhora de até 10% na qualidade.

Em [41] é apresentado um sistema para a avaliação da pronúncia da língua francesa para estudantes que possuem o inglês americano como primeira língua. Nesse trabalho, vários algoritmos que utilizam métricas distintas para pontuação automática foram implementados. Esses algoritmos computam a similaridade entre as pronúncias contidas em um *corpus* de fala de treinamento e, dessa forma, obtêm a pontuação de forma automática. Para a realização dos experimentos, professores franceses atribuíram notas para a pronúncia dos alunos e essas notas foram comparadas àquelas geradas automaticamente pelo sistema. Como resultados, constatou-se que a métrica de número de palavras ditas por período de tempo (*Rate of Speech*) foi a que melhor se correlacionou com as avaliações dos professores. Isso corrobora a hipótese dos autores de que alunos mais avançados no estudo da língua tendem a falar mais rápido que os iniciantes, e isso pode ser um bom indicador de fluência.

No trabalho apresentado em [17] é realizada uma avaliação quantitativa da fluência de estudantes de uma segunda língua através de tecnologias de ASR. Para isso, foi conduzido um experimento similar ao supracitado, onde as avaliações de especialistas foram comparadas àquelas atribuídas automaticamente pelo sistema. A principal diferença entre os dois trabalhos é que no segundo foi utilizada uma gama maior de métricas. Apesar disso, os dois trabalhos apresentaram como conclusão que o número de palavras por período de tempo foi a métrica que mais refletiu a avaliação realizada pelos especialistas.

Em [54], diferente dos trabalhos de [41] e [17], é apresentada uma abordagem de ASR para obter automaticamente precisão do conteúdo de uma resposta falada levando em consideração informações de mais alto nível. A hipótese dos autores na realização desse trabalho foi a de que boas redações se assemelham umas às outras na escolha de palavras, e portanto isso também deveria se aplicar às respostas orais. Na proposta do trabalho, a obtenção da pontuação automática dos testes de fala são obtidos através do cálculo de similaridade entre a transcrição automática de cada teste a ser avaliado e de testes previamente pontuados. Como conclusão, os autores relatam que foi obtida boa correlação entre as avaliações manuais e automáticas quando não haviam erros nas transcrições geradas pelo ASR, caso o contrário, isso não era observado.

Em [7] a compreensão de leitura é apresentada como uma das principais preocupa-

ções das instituições educacionais, pois garante a capacidade dos alunos compreenderem uma fonte de informação e aprender a partir da mesma com precisão. Nesse trabalho, um classificador automatizado é construído para verificar, através da interação pela fala, se um determinado aluno compreendeu ou não a informação nos documentos de estímulo fornecidos.

A diferença do presente trabalho em relação aos trabalhos apresentados anteriormente é a maneira pela qual a avaliação é realizada e o objetivo da mesma. O presente trabalho tem como objetivo o desenvolvimento de uma ferramenta par a avaliação da fluência em larga escala, sendo que a mesma é realizada de maneira automatizada com a possibilidade de intervenção de avaliadores humanos quando essa classificação automática não é possível. Ainda como diferencial do presente trabalho, essa avaliação é realizada a partir de arquivos de áudio contendo a gravação de leituras realizadas por alunos em fase de alfabetização, e tais arquivos devem ser produzidos pelos próprios professores nos mesmos ambientes onde os alunos são alfabetizados, não sendo necessário qualquer tipo de especialização por parte do professor ou mesmo equipamentos especiais para a gravação da leitura.

O uso de técnicas de ASR também já foi explorado na área clínica. Por exemplo, em [14] e [40] a utilização de ASR é investigada na avaliação automática da fala e leitura de crianças com suspeita de possuírem distúrbios decorrentes de doenças que comprometem essas capacidades. Os resultados reportados evidenciam que abordagens ASR são promissoras nessa área, apesar de um conjunto de testes pequeno e ainda faltar uma avaliação mais extensiva dos métodos.

2.2.3 **Estrutura de um ASR**

A tecnologia presente nos sistemas de ASR atualmente é fundamentada em quatro pilares:

1. o **modelo acústico**, responsável pela interpretação do sinal de áudio e sua conversão em palavras e sentenças, associando partes do sinal de voz à unidades de fala (geralmente fonemas) [27];

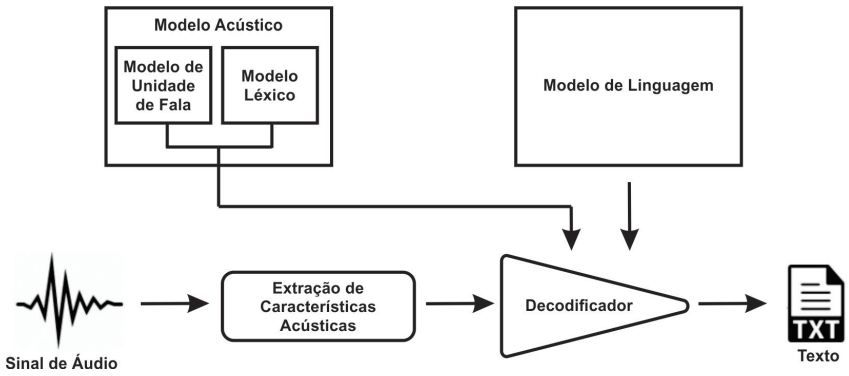
2. o **modelo léxico**, que possui todas as palavras conhecidas pelo ASR e suas transcrições fonéticas, tal como as possibilidades de palavras baseadas em uma sequência fonética (esse modelo faz o mapeamento de sequências de fonemas em palavras utilizando o dicionário presente no sistema) [27];

3. o **decodificador**, que julgando as possíveis hipóteses, é responsável por encontrar a melhor sequência de palavras [27];

4. o **modelo de linguagem**, para caracterizar a língua e gerenciar a combinação de palavras, impedindo que frases agramaticais sejam formadas, atribuindo probabilidades para ocorrência de sequências de palavras [27].

A Figura 1 ilustra a relação entre esses modelos: são extraídas as características acústicas do sinal de áudio e o decodificador gera o texto a partir das mesmas utilizando o modelo acústico e o modelo de linguagem.

Figura 1 – Arquitetura de um sistema ASR



Fonte: Elaboração própria, 2019

Na etapa do **pré-processamento**, o áudio é segmentado em partes com fala e sem fala. Como o ASR usa de probabilidade para classificar as partes do som, ruídos de fundo tendem a interferir no resultado do reconhecimento; desse modo, essa segmentação evita que partes do áudio que não possuem fala interfiram no resultado do reconhecimento de voz. Algoritmos que consideram a energia do sinal de áudio são utilizados de maneira que quando a energia é maior que um valor base, é detectada o início da fala, da mesma maneira que quando a energia é inferior a esse valor, sabe-se que não há fala naquele intervalo [33].

Na **extração de características** são realizadas observações acústicas sobre intervalos de tempos regulares, geralmente de 25ms. Essa etapa permite corrigir algumas características que possam influenciar também no reconhecimento de fala, como por exemplo diferenças nos áudios captados a partir de microfones diferentes. Feito isso, é gerado um vetor com as características extraídas do áudio.

A **decodificação** é o processo de calcular a sequência de palavras que mais se aproxima das características presentes no vetor construído na etapa anterior. Nessa etapa são necessárias as informações de um modelo acústico, um dicionário (tipicamente uma lista de palavras e os fonemas que as representam) e um modelo de linguagem com as possibilidades de palavras e sequência de palavras. Nessa etapa, é fundamental que se saiba quais as palavras podem ocorrer na texto. Essas palavras estão presentes no dicionário

e com o uso de gaussianas obtém-se o vetor com as probabilidades das palavras. A essa altura, o resultado é uma lista de hipóteses para o texto a ser gerado baseado no um sinal de áudio. Na prática, essas hipóteses são as 5 ou 10 mais prováveis. Por fim, entre elas é escolhida aquela cuja a probabilidade seja a maior depois de comparadas com o modelo de linguagem.

Organizado dessa maneira, o sistema pode mapear grande parte da estrutura linguística de um idioma específico para reconhecer o que foi dito.

2.3 DESAFIOS PARA A AUTOMATIZAÇÃO DA AVALIAÇÃO

Inúmeros desafios estão associados ao uso de um ASR, principalmente quando aplicado para avaliar a capacidade de leitura. O processo de reconhecimento de fala apresenta limitações oriundas de questões técnicas que geram dificuldades no processo e podem ser causadas por diferentes motivos, como por exemplo: tamanho do vocabulário, reconhecimento de fala contínua, entre outros. Também existem dificuldades relacionadas à estrutura complexa da voz humana: sotaque, entonação, velocidade da fala, estado emocional etc. [27].

Embora o uso de ASR para avaliação de leitura tem se dado, principalmente, para avaliação do aprendizado de uma segunda língua, estes métodos podem ser utilizados em outros contextos, como por exemplo na avaliação automática da fluência em leitura de crianças em fase de alfabetização de sua primeira língua. Apesar dos desafios encontrados, a automatização traz consigo vantagens, uma vez que ela tende a tornar o processo de avaliação mais ágil, mais barato e com taxas de erro aceitáveis, podendo ser uma forma eficaz de realizar esse tipo de avaliação em larga escala.

Os sistemas para avaliação automática de fluência implementam modelos probabilísticos, o que os tornam passíveis de gerar resultados falso-positivos ou falso-negativos. Erros ocorrerão principalmente quando houver a existência de interferências no áudio que dificultem a avaliação automática, interferindo nas métricas resultantes da avaliação. Nas avaliações em larga escala, por sua vez, é custoso controlar todas as variáveis que possam dificultar a correta captação do áudio (como uso não padronizado de equipamentos de captação, interrupção do professor, sala sem isolamento acústico, etc). Para aumentar a confiabilidade de sistemas para avaliação, é necessário definir parâmetros que minimizem a existência de falso-positivos e falso-negativos.

Existem avaliações experimentais realizadas pelo CAEd/UFJF para a avaliação da oralidade com o objetivo de coleta de dados para a definição de políticas públicas, como o *Mais Alfabetização*[1], o *PAEBES*[2] e o *SAETHE*[3]. Nesses programas, as avaliações são

[1] https://maisalfabetizacao.caeddigital.net
[2] http://www.paebes.caedufjf.net/
[3] http://www.saethe.caedufjf.net/

realizadas de maneira manual para que se possa garantir a confiabilidade dos dados, o que faz com que elas sejam demasiadamente dispendiosas. O uso de sistemas automatizados em conjunto com avaliadores manuais pode aumentar a confiabilidade dos resultados e garantir uma diminuição expressiva dos custos operacionais do processo de avaliação.

3 AVALIA ONLINE: AVALIAÇÃO DA ORALIDADE EM LÍNGUA PORTUGUESA

Este capítulo apresenta o sistema Avalia Online e o seu funcionamento. Na seção 3.1 são apresentadas as Avaliações realizadas pelo CAEd na área da oralidade, seu funcionamento, objetivos e resultados obtidos. A seção 3.2 traz as motivações para a automatização dessas avaliações. A seção 3.3 fala do Sistema Avalia Online e todo seu processo de desenvolvimento. Por fim, a seção 3.4 apresenta a arquitetura do sistema desenvolvido.

3.1 AVALIAÇÕES DE ORALIDADE PROMOVIDAS PELO CAED

O Centro de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora (CAEd/UFJF) é uma instituição que elabora e desenvolve programas destinados a mensurar o rendimento de estudantes das escolas públicas em 15 estados do Brasil, e também atua na formação de especialistas em educação pública e no desenvolvimento de tecnologias de administração escolar. A instituição é uma referência nacional voltada para programas de avaliação educacional.

O CAEd coordena avaliações aplicadas em larga escala, frequentemente realizadas junto a fundações e outras instituições, assim como também realizadas junto a municípios, estados e o próprio Governo Federal. O CAEd usa de fatores intra e extra escolares para a obtenção de medidas de desempenho, tais medidas são utilizadas como base para ações que visam a melhoria da qualidade da educação [8].

As avaliações em fluência em leitura têm como principal objetivo mensurar a qualidade do ensino e encontrar as dificuldades no processo de alfabetização, uma vez que identificadas essas dificuldades, é possível a tomada de decisões para a melhoria do processo.

Na organização das avaliações realizadas pelo CAEd existem diferentes responsabilidades abrangendo todo o processo avaliativo, desde o planejamento e elaboração dos testes até o processamento dos dados resultantes. A realização de uma avaliação tem início com a coleta e processamento de dados para a elaboração dos projetos (como o número de escolas, alunos participantes, região de abrangência, faixa etária dos alunos etc), a partir do processamento desses dados, há a elaboração do material gráfico/visual pra composição de cadernos de testes, assim como a definição dos itens que vão compor as avaliações.

Após a aplicação das provas, existe a coordenação responsável por estudar e publicar os resultados dos projetos por meio de relatórios técnicos. Nessa etapa, o CAEd também ajuda as redes de ensino/parceiros na interpretação dos resultados obtidos nas avaliações.

Para a realização das avaliações, o CAEd estabelece como o conceito de *fluência em leitura*, a capacidade de um leitor **decodificar uma sequência de símbolos e**

pronunciá-los corretamente. De acordo com a sua capacidade de leitura, os alunos são classificados em níveis de fluência: pré-leitor, iniciante ou fluente.

As leituras realizadas pelos alunos compreendem listas de palavras pré-selecionadas na língua portuguesa, lista de pseudopalavras que não existem na língua portuguesa e ainda textos de referência, que são criados pelo CAEd com o intuito de colher informações sobre a pronúncia correta das palavras, compreensão do que está sendo lido e capacidade de interpretar corretamente os sinais de pontuação. Para que as avaliações sejam abrangentes e reflitam a habilidade de leitura de maneira geral e não apenas em um texto específico, os alunos são avaliados a partir de um conjunto de textos. Para as listas de palavras e também para os textos de referência, a escolha das palavras é elaborada buscando levantar informações pertinentes para o relatório final. Na lista de pseudopalavras, estas são criadas de maneira sistemática e levantam informações sobre a assimilação de padrões de pronúncia dentro da língua portuguesa. Todas as leituras têm duração de um minuto, intervalo no qual o aluno deve ler tantas palavras quanto conseguir, e da gravação dessas leituras serão extraídas posteriormente as métricas utilizadas para a avaliação.

Para avaliar a qualidade de leitura, é utilizado o conceito da *precisão*. A precisão na leitura de um texto é calculada como a porcentagem de acertos na leitura, ou seja, quantas palavras foram lidas corretamente **(QPC)** em relação ao total de palavras lidas **(QPL)**. Nas avaliações realizadas pelo CAEd, são classificadas como fluentes aquelas leituras que em um minuto apresentam mais de 65 palavras corretas no texto de referência com um mínimo de 90% na precisão. Aquelas leituras das listas de palavras com 10 palavras ou menos, são classificadas como pré-leituras. As leituras que não forem classificadas como uma das anteriores, são classificadas como leituras iniciantes. Além das métricas já explicadas, na leitura de palavras e pseudopalavras o CAEd extrai outras métricas, como a quantidade de palavras lidas incorretamente e a quantidade de palavras não lidas. Também é avaliado se a criança leu as palavras na ordem correta, se houve identificação de fadiga ocular e se a criança não conseguiu ler nada. Na leitura do texto de referência são coletados a quantidade de palavras lidas corretamente, palavras lidas incorretamente ou mesmo a as palavras não lidas, o total de palavras lidas e a precisão da leitura, além de informações sobre fadiga ocular, seguimento da ordem do texto e a não leitura do texto.

Atualmente essas avaliações são realizadas de maneira manual pelo CAEd acarretando problemas relacionados com o tempo para a obtenção dos resultados que tende a ser grande e ainda o alto custo das avaliações, uma vez que é necessário mão de obra especializada para realizá-las.

3.2 MOTIVAÇÕES PARA A AUTOMATIZAÇÃO DA AVALIAÇÃO

A classificação da fluência baseada em métricas como as descritas anteriormente, pode ser extraída automaticamente das leituras e trás consigo a motivação do desenvolvi-

mento de um sistema voltado para a avaliação da fluência de forma automática. A partir do uso de um sistema de reconhecimento automático de fala (ASR) e de um algoritmo que alinhe os fonemas com o que deveria ser lido no texto de referência, é possível verificar se as palavras foram ou não pronunciadas de maneira correta de acordo com sua transcrição fonética. Dessa maneira, no presente trabalho foi desenvolvido o sistema **Avalia Online**, com o objetivo de auxiliar no processo de correção dos testes de leitura de textos.

Uma vez estabelecidas as métricas para a classificação das leituras, haverá uma fronteira entre os possíveis resultados na avaliação do áudio. Nas proximidades dessa fronteira pode-se identificar uma região de incerteza do sistema, onde pequenas variações nas métricas obtidas podem fazer com que o resultado na avaliação do áudio seja alterado, essas variações podem ser resultado de características do áudio como por exemplo, o aplicador da avaliação falando junto ao leitor, a qualidade do áudio fazer com que ele tenha 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, e aqueles cujas métricas os coloca próximos a essa fronteira, têm classificações que não são tão confiáveis e, por esse motivo, necessitam passar por uma avaliação manual. As avaliações manuais são avaliações cujas métricas e resultados são confiáveis, uma vez que são realizadas por avaliadores capacitados e preparados para esse tipo de avaliação. Dessa maneira, 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, comparando as avaliações automáticas e manuais das mesmas leituras. Sendo assim, aliar o uso do reconhecimento automático de fala com uma abordagem colaborativa, trás ao sistema a robustez da automatização com a confiabilidade desejada, tornando um sistema confiável para a utilização em larga escala e ao mesmo tempo, um sistema economicamente mais viável.

Considerando as possibilidades de erros na automatização do processo, é importante levar em consideração a **precisão** e a **revocação** nos resultados das avaliações realizadas pelo sistema. A precisão do resultado representa a porcentagem de elementos classificados corretamente pelo sistema. A revocação, por sua vez, representa a porção de elementos classificados que de fato é correta, por esse motivo, é um requisito do sistema que o mesmo esteja voltado para uma alta precisão em seus resultados. 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 classes iniciante e pré-leitor do CAEd foram agrupadas em uma mesma classe denominada *não fluente*, uma vez que classificado dessa maneira, o áudio deverá ser encaminhado à avaliação manual.

Considerando que a abrangência das avaliações realizadas pelo CAEd se estende por diferentes regiões do Brasil, é necessário levar em conta as características relacionadas à oralidade, como o caso de sotaques. Os sotaques e outras características linguísticas de cada região influirão na sonoridade da leitura dos textos, sem que isso tenha relação direta com a corretude da leitura. Dessa maneira, é preciso que o sistema possa ser configurado/adaptado de modo que essas características sejam consideradas.

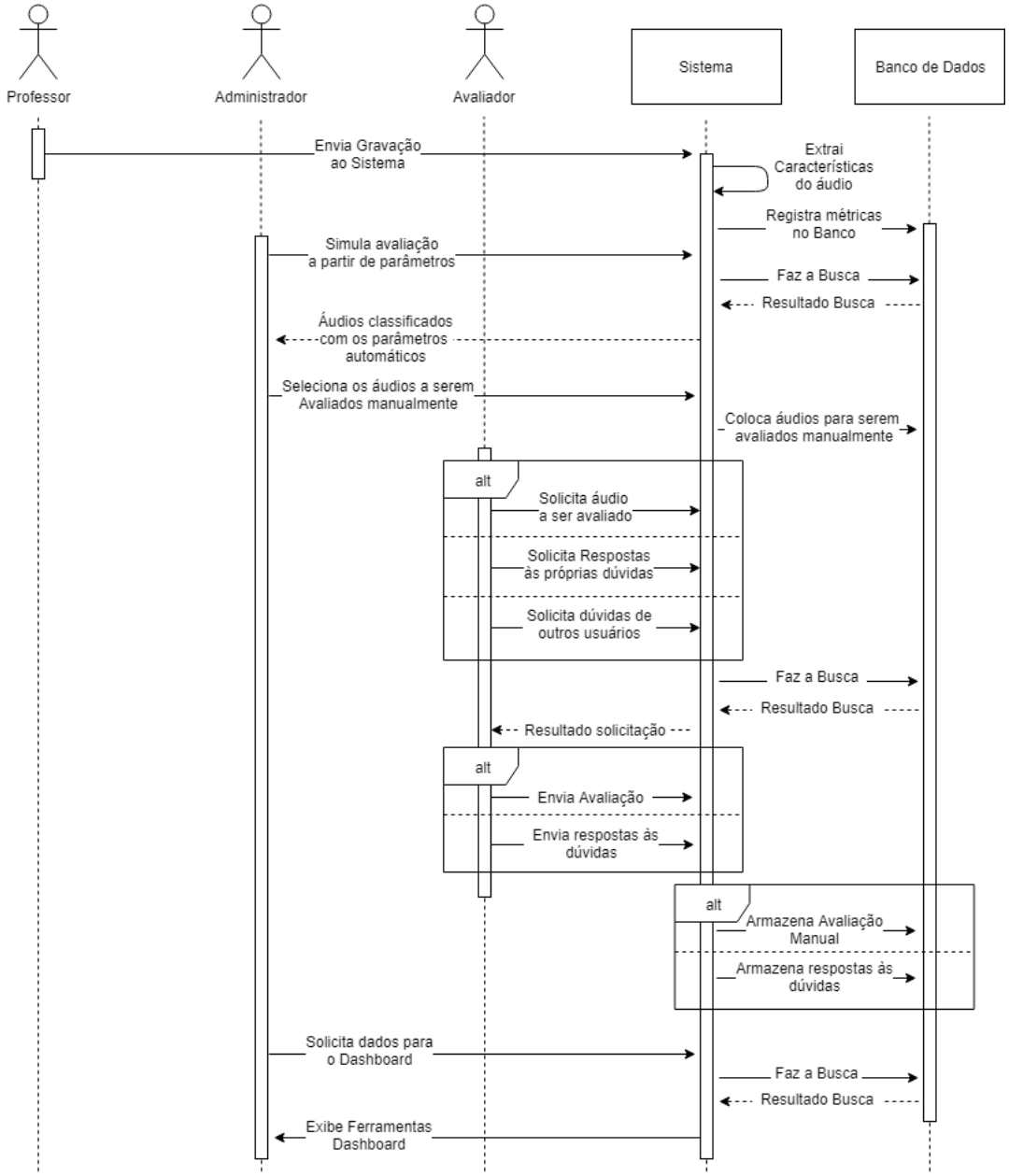
A principal vantagem trazida por uma ferramenta para a automatização de um processo é a otimização do mesmo, resultando em rapidez na coleta dos dados, o uso de grandes amostras, menor custo de administração e taxas de retorno mais expressivas [24]. Assim, com uma ferramenta dessa natureza, os avaliadores 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, o que os ajudará a obter um resultado melhor e mais específico.

3.3 DESENVOLVIMENTO DA SOLUÇÃO

O sistema **Avalia Online** foi desenvolvido para o auxílio no processo de correção dos testes de leitura. Através do sistema, é possível a avaliação da leitura de crianças em fase de alfabetização. De acordo com as métricas extraídas de uma leitura, o sistema pode classificá-la como fluente, ou enviá-la à avaliação manual.

O diagrama de sequência apresentado na figura 2 descreve o funcionamento do sistema.

Figura 2 – Diagrama de Sequência do sistema Avalia Online



Fonte: Elaboração própria, 2019

O professor envia ao sistema os áudios com a gravação das leituras realizadas pelos alunos, o sistema processa esses áudios e extrai suas características, que são armazenadas em um banco de dados. No passo seguinte, para que o sistema possa classificar as leituras de maneira automática, é necessário que hajam parâmetros a serem comparados com as métricas extraídas de cada leitura.

Nessa etapa, o administrador do sistema utilizando-se das métricas extraídas automaticamente junto com as métricas extraídas manualmente de uma porção da base, realiza simulações com com essa porção das avaliações para achar os melhores parâmetros a serem utilizados para a classificação das mesmas. Com os resultados das simulações, o

administrador obtém o melhor conjunto de parâmetros a serem utilizados e os aplica no restante da base. A partir da decisão do administrador do sistema a respeito das métricas para a classificação de um áudio, o sistema tem condições de avaliar automaticamente uma leitura como *Fluente* ou como *Não-fluente* baseando-se nas métricas estabelecidas.

Uma vez que todos os áudios de um projeto sejam classificados segundo os parâmetros escolhidos pelo administrador, o mesmo pode selecionar mais áudios para que sejam avaliados manualmente; em geral, aqueles classificados automaticamente como não fluentes e aqueles cujas métricas os deixam próximos à fronteira entre as possíveis classificações. Os áudios enviados à avaliação manual passam a ficar à disposição dos avaliadores.

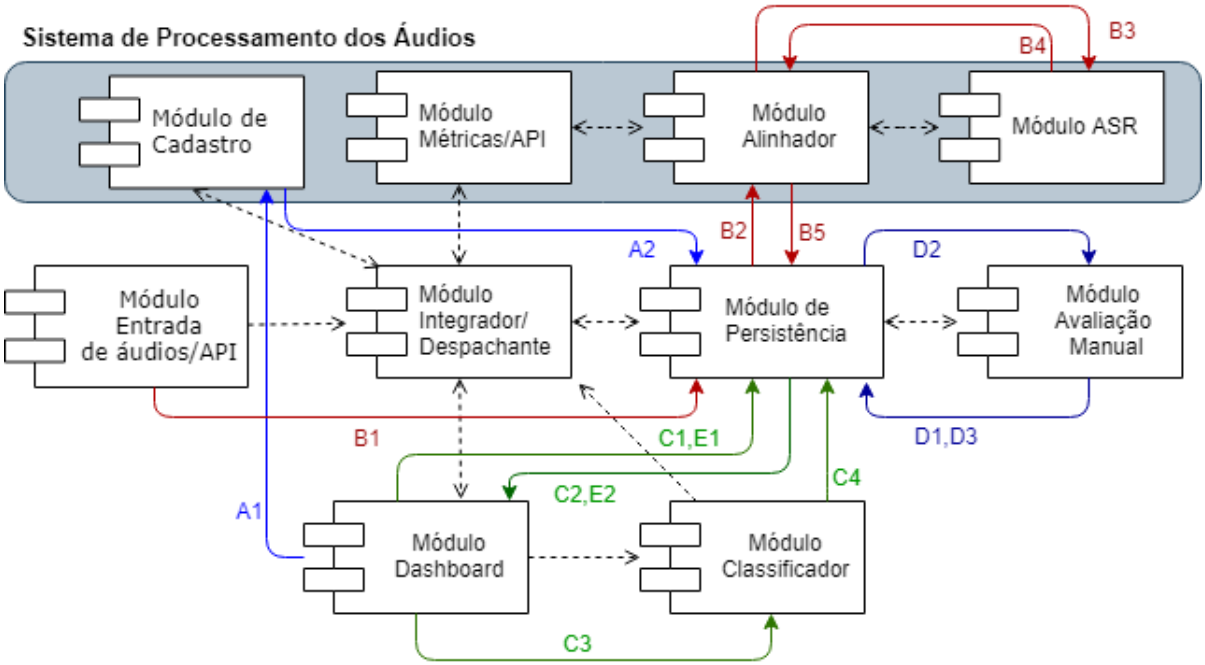
Os avaliadores solicitam ao sistema áudios a serem avaliados e realizam essa avaliação de maneira manual. Caso existam dúvidas durante a avaliação, esses avaliadores podem destacá-las para que sejam auxiliados por outros avaliadores antes de um parecer final sobre as métricas daquele áudio. Para que isso aconteça, além da avaliação de um áudio, também é apresentado ao avaliador as opções de responder as dúvidas de outros avaliadores e de acessar as repostas às suas próprias dúvidas.

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 avaliadores e etc. O objetivo do sistema ao fornecer uma visão geral dos dados é que o administrador possa tomar as decisões para a melhor configuração na avaliação dos áudios.

3.4 ARQUITETURA DO SISTEMA

A figura 3 representa a arquitetura do sistema. O sistema Avalia Online é dividido em módulos que separam suas funcionalidades as deixando independentes, o que favorece a escalabilidade e facilita as manutenções e atualizações. Cada um dos módulos que compõem a ferramenta estão detalhados nas subseções seguintes.

Figura 3 – Arquitetura do sistema Avalia Online



Fonte: Elaboração própria, 2019

As setas identificadas na figura 3 indicam o fluxo de execução do sistema:

A. **Cadastro**

O Módulo *Dashboard* cadastra um projeto [A1] no Sistema de Processamento dos Áudios, e as informações pertinentes do projeto são enviadas ao Módulo de Persistência [A2].

B. **Processamento dos Áudios**

Os áudios são inseridos no sistema através do Módulo de Entrada de Áudios/API, associados a um projeto existente no Sistema de Processamento dos Áudios e enviados ao Módulo de Persistência [B1]. O Módulo Alinhador recebe os áudios e as informações a respeito do projeto a eles associado [B2], e utilizando-se do Módulo ASR realiza a transcrição (é importante ressaltar que a comunicação [B3 e B4] entre os módulos Alinhador e ASR acontece diversas vezes durante a transcrição de um único áudio). Ao final, o Módulo Alinhador retorna ao sistema Avalia Online as métricas extraídas do áudio [B5], que são enviadas ao Módulo de Persistência.

C. **Simulação**

Através do Módulo *Dashboard* é feita uma consulta ao Módulo de Persistência [C1] para obter os dados a respeito das transcrições realizadas. O Módulo *Dashoard* recebe uma resposta [C2] com os áudios cujas métricas estejam de acordo com os parâmetros da consulta e pode então exibir simulações sobre os áudios de um

determinado projeto. As simulações permitem encontrar os parâmetros que melhor se adéquam para a classificação dos áudios de um projeto. Encontrados os parâmetros, os mesmos são usados pelo Módulo Classificador [C3] e realizada a classificação. Em seguida, a classificação é enviada ao Módulo de Persistência [C4], que pode definir quais áudios serão enviados à avaliação manual.

D. **Avaliação Manual**
O Módulo Avaliação Manual requisita ao Módulo de Persistência [D1] um áudio a ser avaliado manualmente. O Módulo de Persistência retorna ao Módulo Avaliação Manual um áudio a ser avaliado [D2]. Realizada a avaliação, o Módulo Avaliação Manual retorna [D3] ao Módulo de Persistência as métricas manuais relacionadas ao áudio avaliado.

E. **Visualização de Resultados**
O Módulo *Dashboard* faz uma requisição ao Módulo de Persistência [E1] a respeito das métricas e dos áudios de um ou mais projetos. O Módulo de Persistência retorna ao Módulo *Dashboard* [E2] os dados solicitados, onde serão exibidos em forma de gráficos, tabelas e outras interfaces que fornecem uma visão generalizada do projeto e permitem a tomadas de decisões.

3.4.1 Sistema de processamento dos áudios

O Sistema de Processamento dos Áudios é um sistema auxiliar, composto pelos seguintes módulos: Módulo de Cadastro, Módulo de Métricas, Módulo Alinhador e Módulo ASR. Esse sistema é responsável pela transcrição do sinal de áudio presente nos arquivos com as gravações das leituras realizadas pelos alunos e pela geração das métricas de cada arquivo. Esse sistema é acessado através de uma interface web. Para que essa utilização aconteça, o primeiro passo é o sistema externo cadastrar um novo projeto . Isso é feito a partir do assistente de criação de projetos, através do Módulo de Cadastro. Durante esse cadastro as informações necessárias são: o texto de referência, o modelo acústico utilizado e as transcrições fonéticas para cada palavra (modelo Léxico). Após completar o cadastro, as informações do novo projeto são retornadas como um arquivo no formato *JSON* para o sistema externo. A partir daí, o sistema externo precisa manter apenas o campo *idExterno* e com ele será possível realizar todos os tipos de requisição ao sistema.

Para a requisição de uma avaliação automática de um áudio, é necessário passar por parâmetro o *idExterno* do projeto ao qual o áudio será relacionado e o próprio áudio a ser avaliado. O sistema recupera as configurações do projeto e através do Módulo Métricas que aciona o Módulo Alinhador. O Módulo Alinhador executa chamadas recorrentes ao módulo ASR e faz o alinhamento do áudio. A saída do alinhamento é então retornada ao Módulo Métricas, processada e as métricas são contabilizadas.

A figura 4 mostra um trecho do arquivo *JSON* que é a saída do Sistema de Processamento dos Áudios. O arquivo é dividido em duas partes: “metrics” e “alignerOutput”, sendo a segunda parte a saída do Módulo Alinhador que será detalhada mais a diante.

Figura 4 – Exemplo do arquivo de saída do Sistema de Processamento dos Áudios


```
{
  "metrics": {
    "result": 0,
    "totalTranscriptedWords": 34,
    "totalDetectedWords": 33,
    "totalUndetectedWords": 1,
    "precision": 0.9705882,
    "detectedUniqueWords": ["abria","as","aves","cada","contava"
      ,"cor","depois","dias","diferentes","dono","dos","e","elas"
      ,"era","ia","manhã","muitos","o","os","outros","pato","patos"
      ,"pela","penas","portão","tamanho","tinha","todas","todos"
      ,"uma","uns","ver","vez"
    ],
    "undetectedUniqueWords": [
      "o"
    ]
  },
  alignerOutput
  }
}
```


Fonte: Avalia Online, 2019

Na parte “metrics” é apresentado o total de palavras transcritas naquele áudio, o total de palavras reconhecidas e de palavras não reconhecidas, o campo precisão representa a divisão das palavras reconhecidas pelo total de palavras transcritas. O próximo campo “detectedUniqueWords” lista todas as palavras presentes no texto e reconhecidas pelo sistema, e o último campo lista as palavras que embora o sistema tenha detectado no áudio o momento que elas deveriam aparecer, as mesmas não puderam ser alinhadas.

3.4.1.1 *Módulo de cadastro*

O Módulo de Cadastro do Sistema de Processamento dos Áudios fornece através de um *pop-up* um assistente para a criação dos projetos. Esse assistente é um *endpoint* específico do Sistema de Processamento dos Áudios para que sistemas externos possam cadastrar seus projetos. Terminado o cadastro do projeto, as informações a respeito do mesmo serão retornadas para o sistema externo, dentre elas, é importante que o sistema externo armazene o **idExterno** para recuperar posteriormente todas as informações do projeto, tal como o texto, o dicionário e as demais propriedades. É relevante notar que a semântica de ***projeto*** no sistema Avalia Online se refere a uma atividade de avaliação da oralidade realizada em crianças em fase de alfabetização, sendo comum que essa avaliação seja realizada com a leitura de diferentes textos. No Sistema de Processamento dos

Áudios, por sua vez, a semântica de ***projeto*** se refere a **um texto**, que será cadastrado no sistema e, a partir daí, o sistema realizará o processamento dos áudios de acordo com as configurações do Modelo Acústico, Modelo Léxico e Modelo de Linguagem realizadas no momento de criação do projeto.

3.4.1.2 *Módulo ASR*

O módulo ASR é responsável por inferir palavras através de uma análise do sinal de áudio. Os sinais de áudio são convertidos em fonemas utilizando o padrão **International Phonetic Alphabet (IPA)** (um conjunto se símbolos utilizados na área de linguística para descrever os sons das linguagens faladas) e as palavras são inferidas com base na sua transcrição fonética e na probabilidade de seguir a palavra anterior de acordo com o Modelo de Linguagem que é utilizado nesse módulo juntamente com os modelos Acústico e Léxico.

A função do Modelo Acústico é a decodificação e extração de características da onda sonora, possibilitando a identificação dos fonemas que ela contém. No Sistema de Processamento dos Áudios, é utilizado um Modelo Acústico treinado com vozes infantis, visto que dessa maneira foram obtidos melhores resultados na extração das características dos áudios.

O Modelo Léxico reúne todas as palavras que o sistema ASR é capaz de reconhecer mapeadas em suas transcrições fonéticas válidas, fornecendo as possibilidades de pronúncia das palavras a partir de uma sequência fonética. Finalmente, o Modelo de Linguagem cria as sequências válidas (ou mais prováveis) de palavras dentro do idioma, procurando evitar frases agramaticais.

O ASR utilizado pelo Sistema de Processamento dos Áudios é o *toolkit* Kaldi[1], uma ferramenta de código aberto destinado a pesquisadores na área de reconhecimento de fala e que tem por objetivo fornecer funcionalidades pertinentes ao desenvolvimento de sistemas para ASR. O *toolkit* Kaldi foi escolhido para ser utilizado pelo Sistema de Processamento dos Áudios por ser uma ferramenta muito utilizada, fazendo com que haja uma grande quantidade de estudos que o utilizam, assim como também diversos pesquisadores. O Kaldi é um sistema estável que continuamente é testado e atualizado, sendo assim uma ferramenta confiável para ser usada no Sistema de Processamento dos Áudios.

3.4.1.3 *Módulo alinhador*

Ao final do processamento do módulo ASR, obtém-se palavras resultantes da associação de um conjunto de fonemas presentes no áudio com a entrada mais próxima no Modelo Léxico, considerando também a palavra mais provável naquela sequência entre

[1] https://github.com/kaldi-asr/kaldi

as palavras possíveis; entretanto isso não é o suficiente para a avaliação da leitura de um texto, uma vez que existe a possibilidade de que as palavras pronunciadas de maneira incorreta não sejam encontradas pela transcrição, ou mesmo o reconhecimento de palavras que não estão presentes no texto. Essas situações são caracterizadas como erros na leitura que não podem ser identificados apenas pelo Módulo ASR. A função do Módulo Alinhador é *forçar* o alinhamento das palavras mesmo quando o áudio possui características que atrapalhem o reconhecimento, como por exemplo uma baixa qualidade de gravação, o som abafado, ruídos externos, interferências do avaliador etc. Essa atuação do Módulo alinhador permite que esses erros sejam detectados pelo sistema.

Para que apenas as palavras de interesse sejam reconhecidas e transcritas nos momentos esperados, existe o **alinhamento forçado**, que alinha temporalmente o que é pronunciado no áudio com o texto de referência. Com o uso do Módulo Alinhador é possível, inclusive, determinar o intervalo de tempo entre início e fim da pronúncia de cada palavra, o que será utilizado na geração das métricas da leitura.

É possível realizar modificações nos modelos para que o alinhamento seja ainda mais preciso. Alterações no Modelo Léxico podem fazer com o que o sistema volte-se especificamente para os fonemas presentes no texto, evitando o reconhecimento de palavras com sonoridade semelhante as do texto mas que não estão presentes na leitura. O Modelo de Linguagem, por sua vez, ao ser alterado pode fazer com que apenas a sequência de palavras presente no texto seja aceita.

O Módulo Alinhador utilizado no Sistema de Processamento dos Áudios é o *Gentle*[2], uma ferramenta de código aberto desenvolvida em *Python* e que realiza o alinhamento forçado recebendo como entrada um áudio e um texto. O Gentle é desenvolvido para utilização em conjunto com o Kaldi, o que facilita sua integração. No Sistema de Processamento dos Áudios a versão padrão do Gentle foi modificada para que atendesse melhor as necessidades do sistema.

A figura 5 exibe um trecho de um arquivo *JSON* que é o resultado do processamento de um áudio realizado pelo Gentle. No início do arquivo o campo “transcript” exibe o texto que foi informado como referência para o alinhamento. O campo “words” armazena uma lista com todas as palavras do texto de referência. Para cada uma dessas palavras, o campo “word” informa qual a palavra no texto de referência, o campo “case” informa a respeito do reconhecimento de alguma pronúncia que seja relacionada àquela palavra. Para o campo “case”, os valores possíveis são “sucess”, significando que algo foi reconhecido no trecho que a palavra deveria ser pronunciada, e “not-found-in-audio” significando que nada foi reconhecido naquele trecho. Os campos “start” e “end” informam os momentos de início e final da pronúncia da palavra. O campo “phones” especifica cada um dos fonemas, com sua duração em segundos. Finalmente, no campo “alignedWord” é exibida

[2] https://github.com/lowerquality/gentle

qual palavra foi alinhada naquele intervalo, caso o valor desse campo seja igual ao do campo “word”, é porque a palavra foi alinhada corretamente, caso contrário, o valor será “unknow”, significando que o que foi reconhecido pelo sistema naquele trecho é diferente daquilo que era esperado.

Figura 5 – Exemplo do arquivo de saída do Gentle


```
"transcript": "penas era uma vez muitos patos cada pato tinha cor e
  tamanho diferentes uns dos outros todos os dias pela manhã o dono ia
  ver as aves depois contava todas elas abria o portão",
  "words": [
    {
      "word": "penas",
      "startOffset": 0,
      "endOffset": 5,
      "case": "success",
      "start": 1.92,
      "end": 2.45,
      "phones": [{"duration": 0.09,"gop": 1,"phone": "p_B"},{"duration":
        0.01,"gop": 1,"phone": "e_I"},{"duration": 0.3,"gop": 1,"phone":
        "n_I"},{"duration": 0.12,"gop": 1,"phone": "æ_I"},{"duration": 0
        .01,"gop": 1,"phone": "s_E"}],
      "alignedWord": "penas"
    },
                                  ⋮
                                  ⋮
    {
      "word": "o",
      "startOffset": 166,
      "endOffset": 167,
      "case": "not-found-in-audio"
    },
    {
      "word": "portão",
      "startOffset": 168,
      "endOffset": 174,
      "case": "success",
      "start": 56.959999,
      "end": 58.289999,
      "phones": [{"duration": 0.08,"gop": 1,"phone": "p_B"}
        ,{"duration": 0.23,"gop": 1,"phone": "o_I"},{"duration": 0.01
        ,"gop": 1,"phone": "ɾ_I"},{"duration": 0.19,"gop": 1,"phone":
        "ə_I"},{"duration": 0.11,"gop": 1,"phone": "t_I"},{"duration":
        0.71,"gop": 1,"phone": "ɐ̃_E"}],
      "alignedWord": "portão"
    }
  ]
}
```


Fonte: Avalia Online, 2019

3.4.2 Módulo API/Entrada de Áudios

O módulo para a entrada dos áudios no sistema Avalia Online é a interface que permite que havendo um projeto, os áudios possam ser carregados e associados ao mesmo. Embora seja importante para o sistema que os áudios das leituras tenham uma qualidade mínima, evitando interferência, ruídos e etc; fundamentalmente, a ideia do sistema é que os áudios sejam captados na própria sala de aula, pelo próprio professor durante as

aulas, não havendo necessidades de equipamentos especiais ou mesmo ambientes voltados especificamente para a captura de áudio. A entrada de áudios do sistema é desenvolvida de maneira que possa ser realizada até mesmo pelo próprio dispositivo utilizado para a gravação do áudio em sala de aula, como por exemplo um *smartphone*.

3.4.3 **Módulo de Persistência**

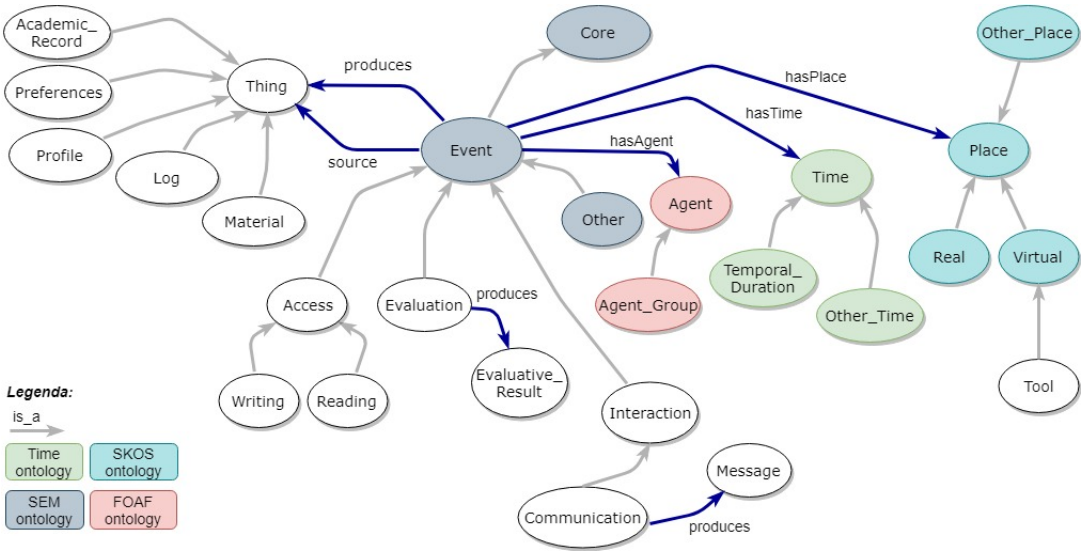
O módulo de persistência de um sistema não provê a persistência dos dados de maneira direta, mas se utiliza de um sistema de gerenciamento de banco de dados (SGBD) auxiliar, por exemplo. No sistema Avalia Online esse módulo provê o acesso ao banco de dados, podendo, inclusive, haver o acesso a mais bancos, caso seja necessário. O banco de dados do sistema Avalia Online armazena todas as informações sobre cada um dos projetos cadastrados no sistema, sobre cada áudio enviado ao sistema juntamente com suas métricas, tanto as obtidas automaticamente quanto as obtidas a partir de uma avaliação manual. Também são armazenadas todas as informações a respeito dos usuários do sistema, tanto os administradores quanto os avaliadores. O banco de dados também armazena os dados sobre os acessos realizados ao sistema e todas as avaliações de leituras realizadas. Todas essas informações são úteis para posteriormente alimentar o Módulo *Dashboard* e através dele fornecer uma visão geral do sistema ao supervisor.

O módulo de persistência pode também exportar os dados armazenados. Como prova de conceito e em consonância com outros experimentos realizados no contexto do LApIC (Laboratório de Aplicações e Inovação da UFJF), foi implementada a exportação no formato *Onto4LA* [13], uma ontologia de eventos criada para auxiliar na interoperabilidade de dados entre sistemas voltados à área de *Learning Analytics*.

A figura 6 representa a ontologia *Onto4LA*. Essa ontologia herda classes das ontologias de alto nível Simple Event Model (SEM)[3], Simple Knowledge Organization System (SKOS)[4], FOAF[5] e Time Ontology[6]. Essas ontologias são utilizadas como ponto de partida para o desenvolvimento da nova ontologia, as classes reutilizadas estão destacadas com cores distintas, conforme a legenda da figura. As setas de cor cinza representam relações do tipo ***is_a***.

[3] https://semanticweb.cs.vu.nl/2009/11/sem/
[4] https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html
[5] http://www.foaf-project.org/
[6] https://www.w3.org/TR/owl-time/

Figura 6 – Visualização da estrutura da Onto4LA



Fonte: Elaboração própria, 2019

Esse trabalho se utiliza dessa ontologia armazenar os dados em um padrão, o que possibilita que os mesmos possam ser interpretados por outros sistemas e utilizados para processamentos de novas informações.

3.4.4 **Módulo Dashboard**

Dentro das aplicações computacionais, os *dashboards* podem ser considerados uma classe específica que nos últimos anos tem sido usada para o desenvolvimento do ensino/aprendizado. Os *dashboards* de maneira geral trazem representações gráficas do estado do sistema juntamente com o histórico de evolução até aquele ponto assim como trazem também informações a respeito de seus usuários (alunos, professores, colaboradores etc.). Essas informações a respeito do sistema, visualizadas de maneira ampla e generalizada, permitem a tomada de decisões voltadas para objetivos específicos [49].

No sistema Avalia Online, 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 realizar simulações. A partir do Módulo *Dashboard* o supervisor também pode acompanhar o andamento das avaliações manuais, gerando gráficos e visualizando dados sobre o trabalho realizado na avaliação manual. As 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. A partir dessas simulações, os supervisores podem definir quantos e quais áudios serão enviados à avaliação manual e também calibrar a classificação automática comparando-a com a classificação manual.

Na figura 7 é apresentada a tela inicial do *dashboard* do sistema. Nessa tela podem

ser visualizados os projetos em andamento, quantos áudios por projeto e dentre esses, a situação na qual se encontram. É exibido quantos áudios já foram processados, quantos não o foram, dentre os processados quantos foram enviados a avaliação manual, e ainda, entre os que já foram avaliados manualmente, quantos os que foram marcados com dúvidas e aguardam opiniões de mais avaliadores.

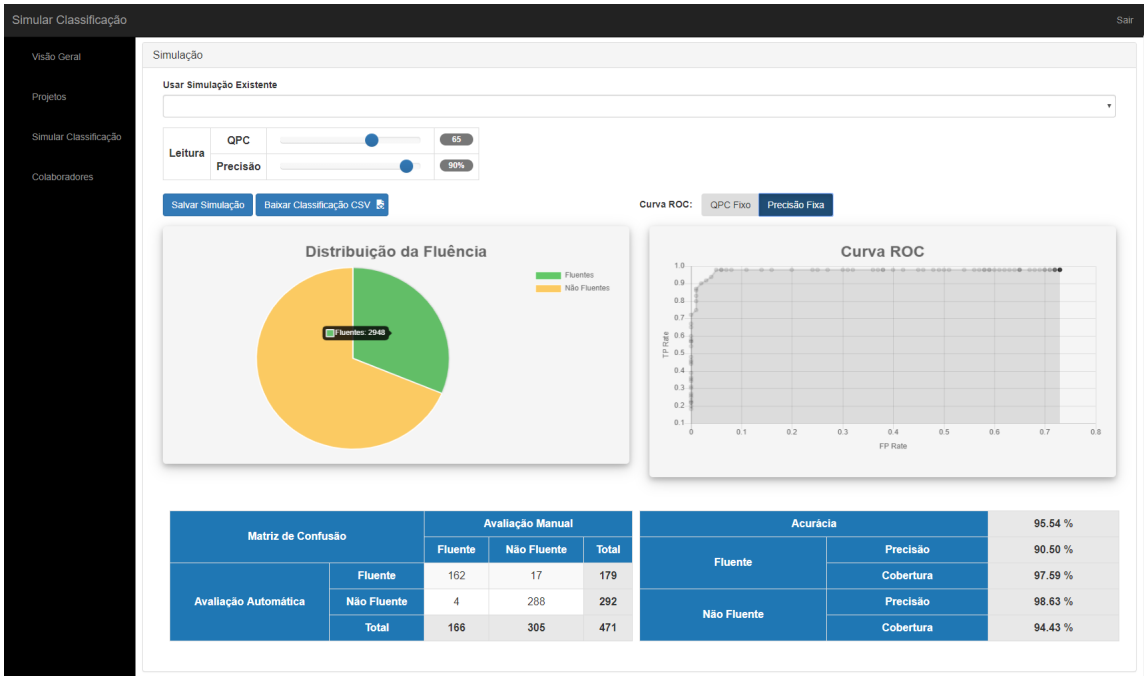
Figura 7 – *Dashboard* do sistema Avalia Online

Fonte: Avalia Online, 2019

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.

O Módulo *Dashboard* possibilita a realização de simulações para a obtenção de métricas e resultados para os projetos cadastrados. A figura 8 apresenta a parte de simulação do *Dashboard*. Ao selecionar valores para o QPC e precisão da leitura, o supervisor obtém um gráfico de pizza da distribuição dos áudios do projeto entre *Fluentes* e *Não-fluentes*. O segundo gráfico é uma curva ROC gerada a partir da variação de valores para o QPC ou para a precisão. A curva representa a análise dessa variação e o impacto que ela gera nos resultados do sistema em termos das taxas de *falsos-positivos* e de *falsos-negativos*, ou seja, nos erros em avaliações automáticas. No eixo horizontal é exibida a taxa de *falsos-positivos*: leituras consideradas fluentes ainda que não o sejam; e no eixo vertical a taxa de *falsos-negativos*: leituras fluentes segundo os parâmetros manuais, mas que avaliadas segundo os parâmetros automáticos, são classificadas como *não-fluentes*.

Figura 8 – Simulação no Dashboard para a base PAEBES



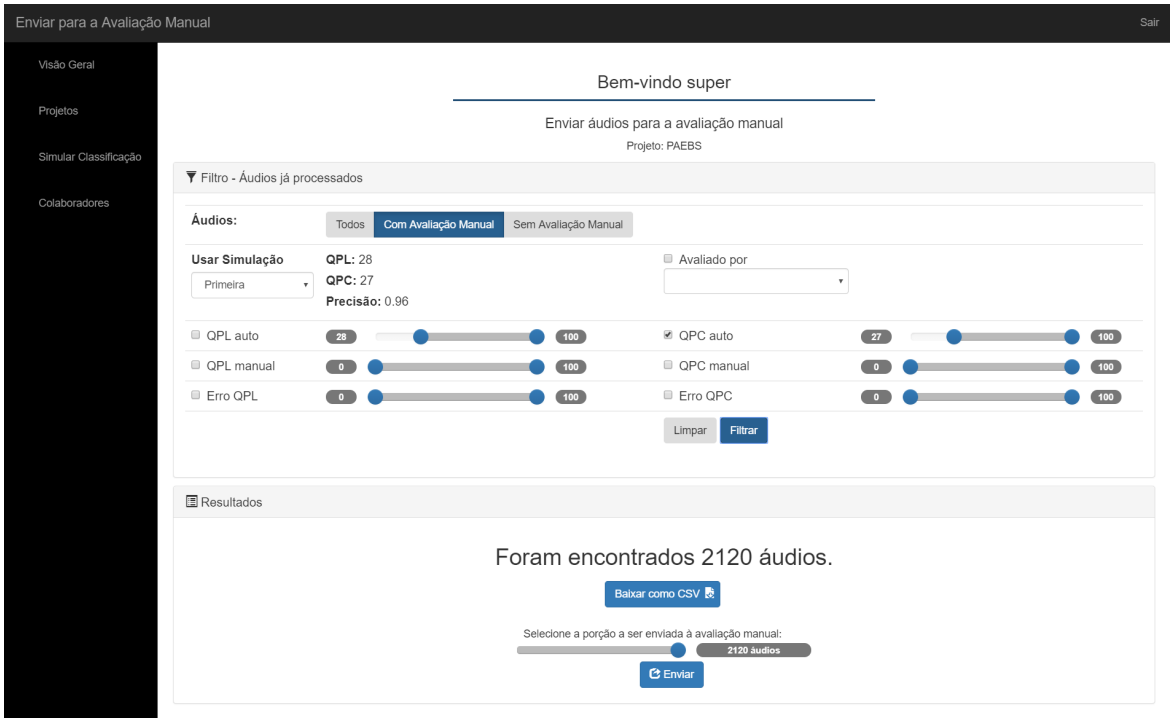
Fonte: Avalia Online, 2019

O terceiro instrumento é a Matriz de Confusão, que 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. Os áudios classificados da mesma maneira tanto nas métricas manuais como nas métricas automáticas representam uma **precisão** na acurácia do sistema, já os áudios que são classificados de maneira errada segundo as métricas automáticas (*falsos-positivos*/*falsos-negativos*) reduzem a cobertura do sistema. Por esse motivo, ao lado da Matriz de confusão, há uma tabela que exibe a acurácia na classificação assim como a precisão e cobertura de cada classe (*Fluente* e *Não-fluente*), essa tabela é atualizada a cada alteração nos parâmetros utilizados.

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. É possível também armazenar no banco de dados os valores de uma simulação pra que sejam utilizados posteriormente, tal como fazer um *download* de um arquivo no formato *'.csv'* com a classificação de todos os áudios da base segundo os parâmetros escolhidos.

O *Módulo Dashboard* possui um filtro exibido na figura 9 que oferece a opção de envio à avaliação manual do áudios relacionados a um projeto.

Figura 9 – Filtragem dos áudios da base PAEBES a serem enviados à avaliação manual



Fonte: Avalia Online, 2019

A partir desse filtro, o supervisor pode enviar à avaliação manual os áudios de um projeto de acordo com valores de alguma característica do áudio ou uma combinação delas. Além de usar valores de simulações já realizadas, é possível filtrar os áudios por faixas de valores de QPL e QPC (automático, manual ou erro). O fitro é aplicado em todos os áudios do projeto e, a partir do resultado, é possível o *download* de um arquivo *'.csv'* que informa quais foram os arquivos filtrados. Também é possível selecionar uma porção dos mesmos para serem enviados à avaliação manual.

3.4.5 **Módulo classificador**

Uma vez realizadas as simulações com parte dos áudios de um projeto, ou mesmo com todo o projeto, o supervisor responsável pode usar das métricas obtidas para a classificação dos áudios daquele projeto. O Módulo Classificador tem como entrada um conjunto de parâmetros e por objetivo a classificação de todos os áudios do projeto. A partir do recebimento de parâmetros e das métricas de um áudio, é possível classificar a leitura como uma leitura satisfatória ou não-satisfatória. A partir do Módulo Classificador são possíveis diferentes classificações baseadas nas variações dos parâmetros.

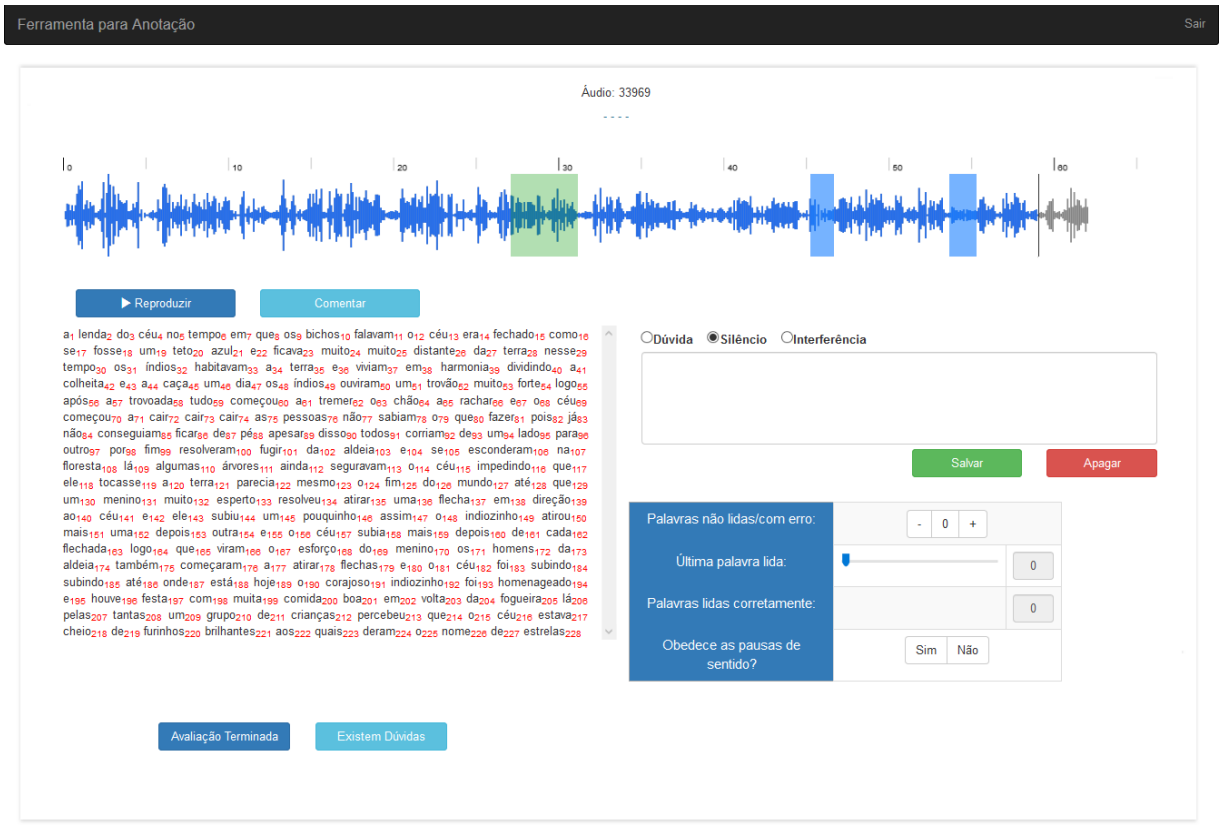
Todos os áudios processados pelo sistema têm as métricas automáticas geradas pelo sistema, os áudios avaliados manualmente tem ainda as métricas geradas a partir de um avaliador humano. Considerando que o sistema é sujeito a falhas e que as avaliações

manuais são realizadas por pessoal treinado, quando um áudio possui ambas as métricas, as métricas automáticas serão desconsideradas e as métricas manuais servirão de referência.

3.4.6 **Módulo avaliação manual**

Esse módulo tem por objetivo possibilitar a avaliação dos áudios por pessoal capacitado, áudios cujas as métricas automáticas os coloquem em uma faixa de incerteza em relação a qualidade da leitura, que foram enviados à avaliação manual pelos supervisores do sistema.

Figura 10 – Interface do Módulo de Avaliação Manual



Fonte: Avalia Online, 2019

No Módulo Avaliação manual, os colaboradores ao avaliar um áudio poderão encontrar os motivos pelos quais foram enviados à avaliação manual. A figura 10 exibe a interface na qual é realizada essa avaliação. Nessa tela o avaliador tem uma representação gráfica da onda sonora gerada pelo áudio referente à leitura realizada, e pode ouvir a leitura (quantas vezes julgar necessário), ou ouvir trechos específicos da mesma enquanto acompanha o texto a que ela se refere. A interface permite ao avaliador destacar trechos no espectro do áudio avaliado e fazer marcações nesses trechos, essas marcações podem representar características específicas daquele áudio (como por exemplo: interferências, erros de pronúncia, mais pessoas falando durante a leitura, trechos sem leitura e etc).

Também é possível representar dúvidas pessoais sobre a extração das métricas pra que o avaliador possa ser auxiliado por outros avaliadores. O fato de haver uma ferramenta específica com as opções propícias às avaliações dos áudios, gera uma maior padronização dos resultados e trás facilidades que não existem quando a avaliação é feita por meio de um *player* de áudio comum. Se ao final da avaliação manual não houver no áudio avaliado qualquer marcação relacionada à dúvidas do avaliador, o avaliador deve preencher os campos relativos às métricas da leitura, que são a quantidade de palavras lidas com erro, a ultima palavra lida, e se a leitura respeita a pontuação obedecendo as pausas de sentido; em seguida, a avaliação deve ser encerrada. Caso existam marcações relacionadas a dúvidas durante a avaliação, essa situação deve ser assinalada e esse áudio será colocado à disposição de outros avaliadores, que ajudarão nas decisões a respeito dessas dúvidas.

A outra parte da ferramenta para avaliação manual, permite ao avaliador acessar trechos de avaliações que foram marcados como dúvidas, tanto os trechos marcados por outros avaliadores, quanto os trechos relativos às próprias dúvidas. Nessa parte da ferramenta, os avaliadores interagem através de troca de mensagens para chegar em um consenso quanto às métricas da leitura naquele áudio. Dessa maneira, os colaboradores aumentam a confiabilidade do processo de avaliação gerando melhores resultados, uma vez que são resultados vindos da discussão entre diferentes avaliadores. Além disso, como a abordagem colaborativa ainda envolve a construção de conhecimento, à medida que os colaboradores interagem entre si, é esperado que os mesmos tenham menos dúvidas e gerem avaliações mais precisas [28].

Todos os módulos do sistema se comunicam através do Módulo Integrador, permitindo que haja a troca de informações necessária. O Módulo Integrador do sistema Avalia Online é desenvolvido em *PHP*, a mesma linguagem dos demais módulos. A comunicação entre os módulos é realizada através de requisições *AJAX* (Asynchronous JavaScript e XML), que é o uso do objeto XMLHttpRequest para se comunicar com os *scripts* do lado do servidor. Através dessas requisições é possível enviar e receber informações em uma variedade de formatos, no sistema Avalia Online é utilizado o formato *JSON* para a comunicação entre os módulos.

4 EXPERIMENTOS

No capítulo 4 são apresentados dois experimentos: o primeiro, detalhado na seção 4.1, tem por objetivo aferir a qualidade das métricas extraídas automaticamente das leituras realizadas. O segundo experimento, detalhado na seção 4.2, avalia a abordagem colaborativa do Módulo de Avaliação Manual e tem por objetivo ressaltar as vantagens da utilização dessa abordagem a partir da comparação com avaliações manuais realizadas da maneira convencional.

4.1 AVALIAÇÃO DO CLASSIFICADOR

O primeiro experimento tem como objetivo aferir, por meio de simulações, a qualidade da avaliação automática daqueles áudios que já possuem avaliação manual.

Nesse experimento foram utilizadas bases compostas por áudios contendo a gravação de leituras realizadas por crianças em fase de alfabetização. Tais bases são fruto de avaliações realizadas pelo CAEd em 4 estados: Espírito Santo, Paraíba, Pará e Pernambuco, contendo um montante de 17054 áudios das bases *PAEBES* (9412 áudios), *Piloto1* (5953 áudios) e *Piloto2* (1689 áudios). A base *PAEBES* possui todas as leituras relacionadas a um único texto, enquanto as demais estão divididas em 9 textos, dessa maneira, a base *PAEBES* tem todas as suas leituras cadastradas em um único projeto no Sistema de Processamento dos Áudios e as demais são dividas em 9 projetos, o que faz com que o projeto que contém a base *PAEBES* seja consideravelmente maior e, dessa maneira, os resultados relacionados a esse projeto sejam mais próximos à realidade.

Os resultados desse experimento são exibidos ao longo do texto, especialmente com a apresentação e análise dos gráficos que são gerados pela ferramenta à medida da realização do experimento.

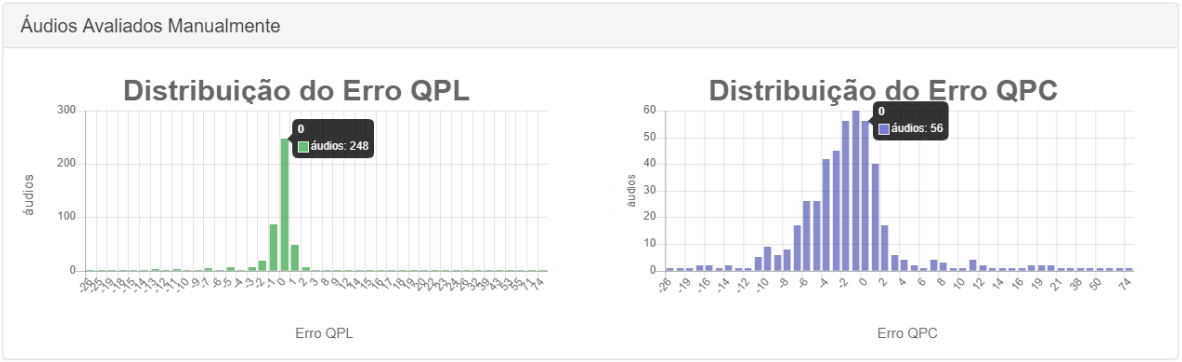
Nesse experimento é importante ressaltar as diferenças entre as **métricas brutas** (erros de QPL e QPC) da **classificação dos áudios**. Nas métricas brutas, o objetivo é definir de maneira automática a *quantidade* de palavras lidas e de palavras lidas corretamente em um áudio. Esses valores servirão para o objetivo da avaliação que é definir quais leituras são fluentes ou não. Assim, ainda que haja um erro nas métricas brutas, o sistema pode definir de maneira correta a respeito da *classificação* da leitura contida em um áudio. O *Dashboard* oferece os instrumentos para as simulações a serem realizadas. Esses instrumentos são úteis para que o supervisor do projeto encontre as métricas que melhor classificam os áudios e devem ser utilizadas para a classificação de toda a base. Essas simulações podem ser realizadas diversas vezes em busca de um conjunto de parâmetros que gere a maior acurácia para o conjunto de áudios avaliados.

No início do experimento, cada uma das bases foi cadastrada como um projeto

diferente no sistema e todos os áudios processados para a extração das métricas automáticas. Em seguida, a partir do Módulo *Dashboard* foram selecionados aleatoriamente 5% dos áudios de cada projeto para serem enviados à avaliação manual.

A figura 11 mostra o gráfico da distribuição de erros na base *PAEBES* exibido no *Dashboard*. Esses gráficos levam em consideração apenas os áudios que possuem tanto as avaliações automáticas como também as avaliações manuais. É possível notar no gráfico que a maioria dos áudios têm Erro QPL (Quantidade de Palavras Lidas) igual a 0, o que indica que na maioria das vezes o sistema consegue extrair de maneira correta a quantidade de palavras lidas naqueles áudios. Nesse gráfico os valores positivos mostram a quantidade de áudios onde o número de palavras lidas é maior que o detectado pelo sistema, enquanto os valores negativos mostram a quantidade de áudios onde o número de palavras detectadas pelo sistema é maior que a quantidade de palavras realmente lidas, o que pode acontecer quando, por exemplo, o aluno hesita ao ler uma palavra e ao repeti-la, o sistema a contabiliza novamente.

Figura 11 – Distribuição dos erros QPL e QPC entre os áudios avaliados manualmente na base PAEBES

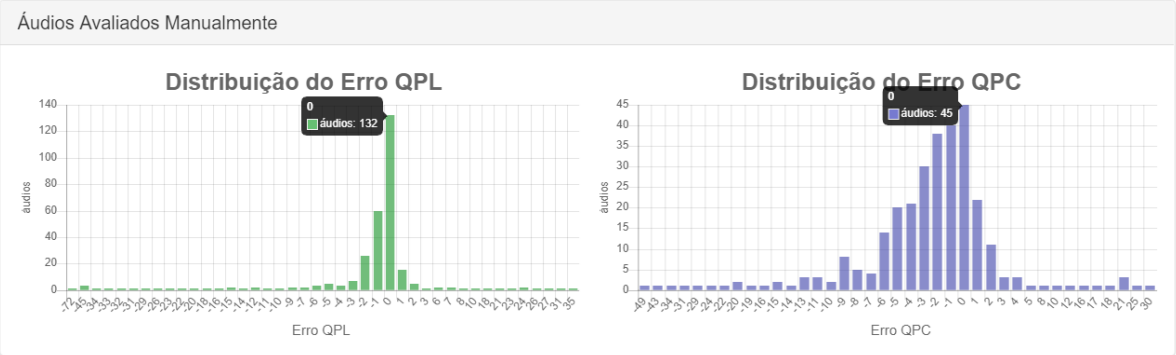


Fonte: Avalia Online, 2019

No gráfico da distribuição de Erro QPC, embora a maioria dos áudios não tenha Erro QPC = 0, ainda se concentra em torno do 0, o que mostra uma sensibilidade maior do sistema em relação às palavras lidas corretamente. Nesse gráfico, os valores positivos mostram áudios nos quais os sistema considerou como lidas corretamente uma quantidade menor de palavras do que as que de fato o foram, e os valores negativos mostram o volume de áudios nos quais o sistema considerou como correta a leitura de uma quantidade de palavras maior do que a quantidade que de fato foi lida de maneira correta, o que pode acontecer quando o erro na leitura de uma palavra for muito sutil e a leitura da palavra, ainda que errada, seja considerada correta pelo sistema. Essas anomalias podem ocorrer quando as sonoridades das pronúncias (incorreta e correta) de uma palavra são parecidas.

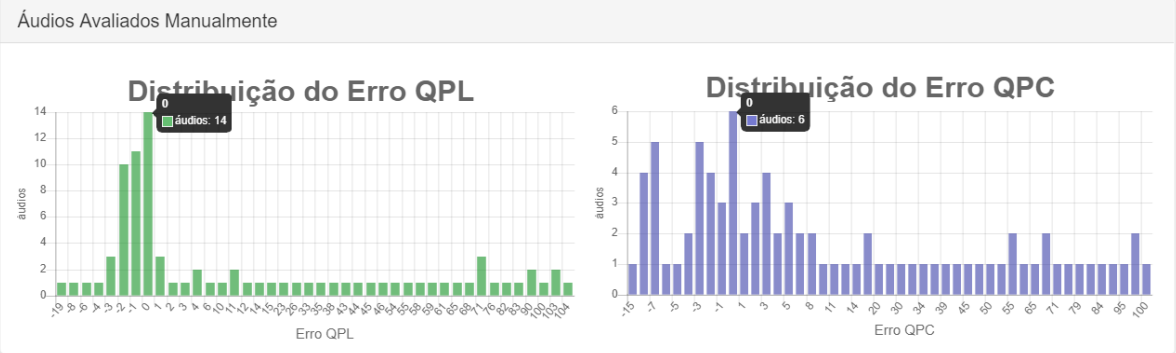
Nas figuras 12 e 13 existe a mesma análise para as bases *Piloto1* e *Piloto2* respectivamente.

Figura 12 – Distribuição dos erros QPL e QPC entre os áudios avaliados manualmente na base Piloto1



Fonte: Avalia Online, 2019

Figura 13 – Distribuição dos erros QPL e QPC entre os áudios avaliados manualmente na base Piloto2



Fonte: Avalia Online, 2019

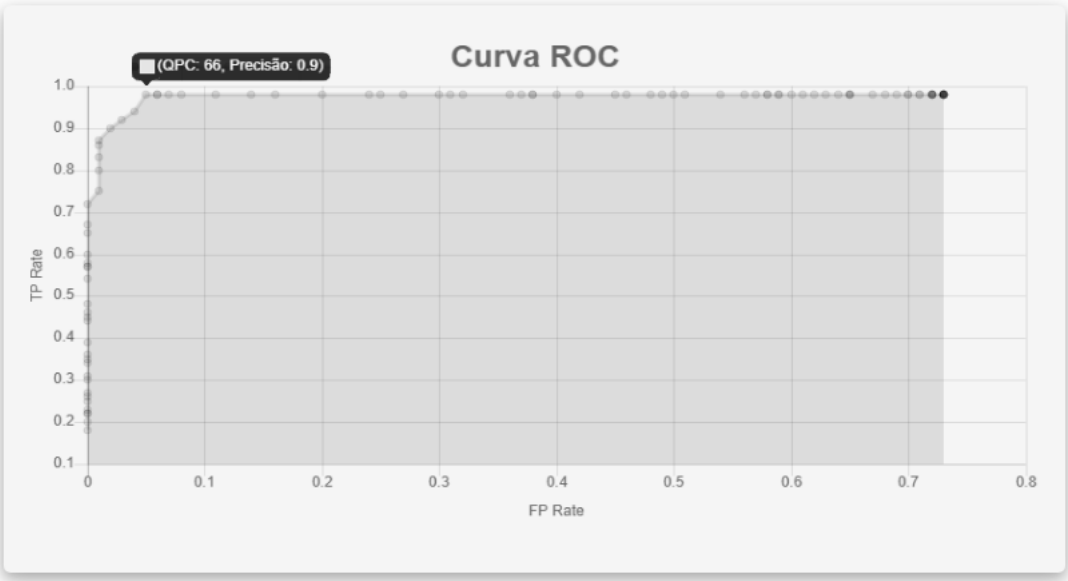
Nas avaliações realizadas pelo CAEd, as métricas utilizadas para a classificação são $QPC = 65$ e $Precisão = 90\%$. A base *PAEBES* classificada com esses parâmetros, apresenta um resultado de 2915 áudios (30,97%) fluentes e 6497 áudios (69,03%) não-fluentes.

Após a realização das simulações na base *PAEBES* com 5% dos áudios avaliados manualmente e usando as métricas utilizadas pelo CAEd, o resultado para toda a base foi 2948 áudios (31,32%) fluentes e 6464 áudios (68,68%) não-fluentes. Dentre os áudios que também possuíam classificação manual, através da Matriz de Confusão 450 áudios (95,54%) foram classificados da mesma maneira tanto com os parâmetros manuais quanto com os parâmetros automáticos, o que é apresentado como a **acurácia do sistema**.

A curva ROC gerada nessa simulação é exibida na figura 14, e seguindo a metodologia apresentada em [12], foi escolhido o ponto da curva com a maior taxa de *verdadeiros-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 *PAEBES* com 5% de seus áudios avaliados manualmente, apresenta para *Precisão = 90%* um *QPC = 66* como o melhor valor para a classificação.

Figura 14 – Gráfico da Curva ROC para a base PAEBES com 5% dos áudios processados



Fonte: Avalia Online, 2019

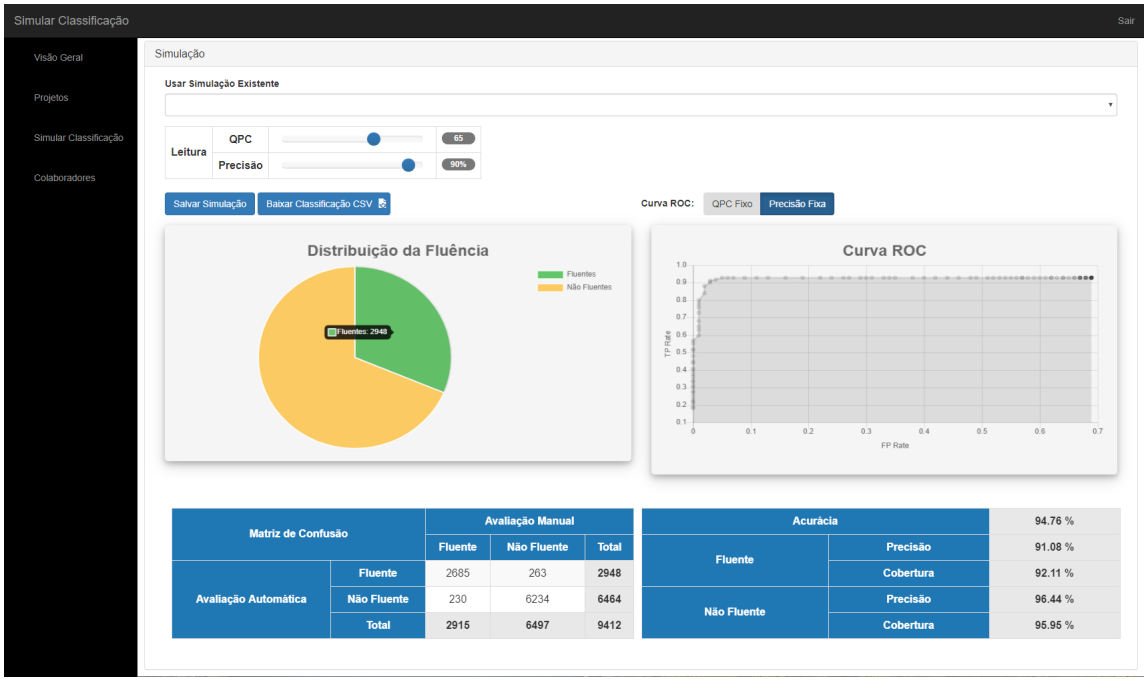
Ao utilizar os valores encontrados a partir da curva ROC, a acurácia do sistema passou a 95,97%, indo de 450 para 452 áudios classificados automaticamente de maneira correta.

A partir do filtro do *dashboard*, mais 10% da base PAEBES foi enviada à avaliação manual.

Com as novas avaliações, a acurácia do sistema passou a 95,75%, enquanto a curva ROC continuou apresentando o *QPC = 66* como o melhor valor para *Precisão = 90%*.

Por fim, o restante da base PAEBES foi enviada à avaliação manual, e os gráficos do *Dashboard* foram gerados novamente com os novos parâmetros e os resultados obtidos são apresentados na figura 15.

Figura 15 – Simulação no Dashboard para a base PAEBES com todos os áudios avaliados manualmente



Fonte: Avalia Online, 2019

Com toda a base avaliada manualmente, com o uso do QPC = 66 encontrado na análise da curva ROC e uma precisão de 90%, foi obtida uma acurácia de 94,99% para toda a base. Assim, mostra-se que o sistema fornece um resultado consistente para avaliar um conjunto grande de dados e que, através das ferramentas fornecidas na interface, é possível ajustar os parâmetros da classificação para alcançar altos valores de acurácia.

A tabela 1 mostra uma comparação entre os valores de acurácia alcançados mantendo-se uma precisão de 90% e sendo realizadas variações de QPC e da porção da base avaliada manualmente.

Tabela 1 – Comparação da acurácia com a variação do QPC para cada porção da base PAEBES avaliada manualmente.

Porção da base avaliada manualmente	**QPC**	**Acurácia**
5%	65	95,54%
5%	66	95,97%
15%	66	95,75%
100%	66	94,99%

Fonte: Elaboração própria, 2019

Os experimentos realizados com áudios gravados por professores de escolas públicas mostraram que a abordagem é capaz de reduzir o trabalho manual, tendo, na base utilizada

nos experimentos, classificado aproximadamente 43% dos áudios como leituras satisfatórias ou insatisfatórias. Caso essa proporção se confirme em uma avaliação em escala nacional, tem-se a redução de mais de 40% do custo com contratação de avaliadores.

Segundo os dados disponíveis no portal do INEP[1] a respeito da Provinha Brasil, no ano de 2018 a prova foi realizada por 5.201.730 alunos, e considerando que esses alunos realizassem a avaliação da oralidade, na qual cada aluno grava 3 áudios de um minuto, seriam gravados 15.605.190 áudios para essas avaliações, resultando em aproximadamente 260 mil horas de gravações. Considerando o resultado do experimento, isso faria com que aproximadamente 112 mil horas de leituras fossem classificadas automaticamente após o processamento dos áudios, evitando que fossem necessárias essas horas de trabalho humano.

4.2 AVALIAÇÃO DA USABILIDADE DO SISTEMA

A ferramenta para a avaliação manual do sistema Avalia Online foi desenvolvida para ser disponibilizada via internet, pra que dessa maneira, os avaliadores possam realizar o trabalho à distância, no local e horário que lhes for mais conveniente, o que gera mais liberdade e autonomia para a realização do trabalho [10].

O segundo experimento é voltado para o Módulo Avaliação Manual e sua abordagem colaborativa, e tem como objetivo comparar as avaliações manuais de um mesmo conjunto de áudios realizadas através da ferramenta e da maneira convencional. Para esse experimento publicado em [11], foi utilizada uma base reduzida (471 áudios) que contém as mesmas características das bases utilizadas no experimento anterior. Depois de criado um projeto, realizado o *upload* dos áudios, as simulações e o processamento, 201 áudios (42,7%) foram classificados de maneira automática e os 270 áudios (57,3%) restantes foram enviados à avaliação manual.

Os resultados do experimento comparam o tempo médio de avaliação dos áudios avaliados tanto através da ferramenta quanto apenas com o uso de planilhas eletrônicas, tal como também leva em conta as considerações realizadas pelos participantes de cada um dos grupos.

A avaliação manual foi realizada por dois grupos distintos com 7 avaliadores cada. O **grupo A** utilizou a ferramenta para avaliação e o **grupo B** a realizou da maneira convencional através de planilhas eletrônicas. Os grupos realizaram a tarefa em momentos distintos. Todos os colaboradores participaram de um treinamento sobre o protocolo de avaliação com os critérios utilizados pelo CAEd para a classificação dos áudios, demonstradas algumas das situações comumente encontradas durante a avaliação e a forma que deveriam agir diante cada uma delas. As avaliações tiveram duração de 1

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

hora.

Durante a avaliação, deveriam ser contabilizadas as palavras lidas incorretamente tal como o quantidade total de palavras lidas. Em caso de dúvida em alguma situação não prevista no protocolo ou que o avaliador não tivesse certeza sobre a decisão a ser tomada, este deveria especificar o trecho contendo a dúvida. Os participantes do **grupo A** deveriam marcar os trechos com dúvidas através da ferramenta. Os participantes do **grupo B** deveriam especificar o intervalo do áudio que gerou dúvida na própria planilha. No **grupo A**, a dúvida seria exposta aos demais colaboradores através do sistema para que ajudassem na tomada de decisão; no **grupo B**, a dúvida foi registrada para que fosse verificado posteriormente por um supervisor, mas o avaliador decidia sozinho quanto as métricas daquele áudio.

Durante o período do experimento, o **grupo A** avaliou um total de 133 áudios, enquanto o **grupo B** avaliou 112. O tempo médio gasto para avaliação de cada áudio foi de 3'21" no **grupo A** e 3'48" no **grupo B**.

A figura 16 exibe a área da ferramenta voltada para a resolução das dúvidas geradas durante a avaliação manual do áudio de uma leitura. Ao acessá-la, o avaliador tem duas abas, a primeira para as próprias dúvidas em aberto e na segunda as dúvidas de outros avaliadores.

Figura 16 – Interface para a abordagem colaborativa

Fonte: Avalia Online, 2019

Cada participante tinha acesso ao trecho dos áudios marcados como suas dúvidas ou as de outros avaliadores, poderia ouvi-lo e, sentindo-se à vontade, responder com a sua interpretação daquele trecho, justificando-se com base no protocolo.

Entre os áudios avaliados pelo **grupo A**, 42 tiveram trechos que foram sinalizados

com algum tipo de dúvida por parte do avaliador, dentre as quais 37 foram respondidas por outros avaliadores com discussões que, no total, geraram 146 respostas. Ao final do experimento, 40 áudios com dúvidas tiveram suas métricas definidas por seus avaliadores, apenas 5 dúvidas não foram respondidas por conta do término da sessão.

Em comparação com a avaliação sem a ferramenta, houveram 12 áudios os quais as métricas definidas em cada grupo os classificava de maneira diferente segundo os parâmetros posteriormente utilizados, esses áudios tiveram marcações de dúvidas ao serem avaliados pelo **grupo A**. Como a decisão sobre o resultado foi discutida entre os colaboradores do **grupo A**, nessa abordagem o resultado tende a ser mais confiável. Percebe-se, também, que a média de dúvidas por avaliador é maior quando utilizada a ferramenta, o que mostra que quando a abordagem é colaborativa, há uma tendência maior de expor as dúvidas e, dessa maneira, uma chance menor de decisões precipitadas.

A tabela 2 apresenta as considerações a repeito das avaliações realizadas por cada grupo.

Tabela 2 – Comparação das avaliações manuais realizadas através da ferramenta e da maneira convencional

Grupo	**Avaliação**	**Áudios**	**Tempo Médio Avaliação**	**Considerações**
grupo A	Sistema Avalia Online	133	3'21"	Dúvidas em 42 áudios 37 dúvidas respondidas Discussões com 146 respostas
grupo B	Planilha Eletrônica	112	3'48"	Trechos com dúvidas ignorados pelos participantes 62,5% dos participantes afirmaram a possibilidade de decisões diferentes caso auxiliados

Fonte: Elaboração própria, 2019

Ao final do experimento, os participantes dos dois grupos preencheram um questionário com perguntas sobre o trabalho realizado. Os participantes do **grupo B** foram questionados sobre critério que utilizaram para dar um parecer final sobre as métricas nos casos em que haviam dúvidas. Em geral, responderam que desconsideraram o trecho do áudio com a dúvida. Dessa maneira são contabilizadas menos palavras lidas do texto, o que pode interferir diretamente na classificação da leitura.

Por fim, 62.5% dos participantes do **grupo B** responderam que, se pudessem ser auxiliados por outros avaliadores em suas dúvidas, é possível que suas decisões quanto às

métricas fossem diferentes, o que mostra insegurança quanto as suas decisões.

Os participantes do **grupo A**, por sua vez, foram questionados se a troca de informações com outros avaliadores fez diferença nos resultados de suas avaliações. Foram registradas 87.5% das respostas afirmativas a essa pergunta, o que mostra que o uso do sistema pode ter auxiliado na construção do conhecimento dos avaliadores e, principalmente, em garantir uma avaliação mais correta desses áudios.

Após o processo de avaliação manual, a maioria dos avaliadores considerou que a troca de mensagens com outros avaliadores gerou uma reavaliação da leitura.

Considerando novamente os dados do portal do INEP a respeito da Provinha Brasil no ano de 2018, e também que cada aluno gravaria 3 áudios para a avaliação da oralidade, e que através da ferramenta Avalia Online teríamos a classificação automática de aproximadamente 43% da base, restariam ainda por volta de 10.455.477 áudios a serem avaliados de maneira manual. Com os resultados desse segundo experimento, sem a ferramenta Avalia Online seria necessário aproximadamente 662.180 horas de trabalho para a avaliação manual de toda a base, enquanto que, através da ferramenta, além de uma maior confiabilidade dos resultados apresentados, as avaliações manuais poderiam ser realizadas num tempo de 583.764 horas de trabalho, gerando assim uma diferença significativa no tempo necessário para a avaliação.

5 CONCLUSÕES

Nesse trabalho foi apresentado o sistema **Avalia Online**, desenvolvido com o objetivo de possibilitar a avaliação em larga escala da oralidade de crianças em fase de alfabetização. Para validar a eficácia do sistema, foram realizados dois experimentos: o primeiro voltado para a classificação automática de leituras quanto à sua fluência, que mostrou a possibilidade da automatização de grande parte do trabalho e sua eficiência e acurácia na classificação das leituras. O segundo experimento, voltado para a usabilidade do sistema, mostrou as vantagens obtidas pela adoção de uma abordagem colaborativa. A partir desses experimentos, podemos concluir que através do sistema é possível reduzir os custos e tempo necessários para a avaliação da oralidade infantil e que a partir da abordagem colaborativa é possível aumentar a confiabilidade dos resultados obtidos. Dessa forma, o sistema permite uma avaliação da oralidade que seja mais abrangente, mais rápida e economicamente mais viável.

No desenvolvimento do sistema foram usadas técnicas de ASR para a extração de métricas em arquivos de áudios contendo leituras realizadas por crianças em fase de alfabetização, e a partir dessas métricas, essas leituras são classificadas. As classificações são baseadas em parâmetros determinados a partir de resultados exibidos em um *dashboard*, que levam em conta as avaliações manuais de uma pequena parte das leituras.

5.1 CONTRIBUIÇÕES

Ao final da realização desse trabalho, é necessário destacar as seguintes contribuições:

1. Inovação do sistema: até onde conhecemos, essa é a primeira solução voltada para avaliações em grande escala no território brasileiro com o objetivo de aferir e classificar automaticamente a fluência na fala de crianças durante seu processo de alfabetização.

2. Eficiência e eficácia: a abordagem permite a entrega de resultados com alta acurácia em um curto intervalo de tempo, além de possuir um baixo custo operacional, isso reduz drasticamente a necessidade de contratação de mão-de-obra especializada para avaliação dos áudios.

3. Modularidade: a solução é modular, podendo ser adaptada para outros tipos de modelos e avaliações. Além disso, é possível adaptar novas instâncias de módulos para adicionar novos comportamentos ou paralelizar o processamento de acordo com a infraestrutura disponível.

4. Reprodutibilidade: Ao garantir um ambiente centralizado para processamento dos áudios, experimentação e geração de relatórios e ainda ferramentas para a avaliação

desses resultados, 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.

5.2 LIMITAÇÕES

Uma vez que o sistema se utiliza do reconhecimento automático de fala, ele traz consigo as limitações relacionadas a essa técnica. Como o sinal de entrada é um áudio contendo a fala do aluno, todas as interferências presentes nesse sinal (tal como ruídos, baixa qualidade da gravação, mais pessoas falando ao mesmo tempo etc) são fatores que limitam a capacidade do sistema. Dessa maneira, a qualidade dos resultados fica reduzida à qualidade da gravação dos áudios.

Outra limitação é a disponibilidade de bases para o treinamento do sistema, que, embora existam técnicas que elevem a qualidade do sistema treinado com vozes de adultos, obteria uma qualidade maior caso houvesse bases suficientemente grandes para treinamento com vozes de crianças da mesma faixa etária daquelas as quais serão avaliadas pelo sistema.

Os testes realizados com o sistema tiveram o foco voltado em seu funcionamento, mas o volume da amostra ainda é muito pequeno em relação a uma cobertura nacional tal como é proposto, desse modo, os resultados mostram que a abordagem é eficaz, embora os valores de acurácia possam sofrer alterações a medida que sejam utilizadas bases maiores.

Não foram realizados testes de *stress* da ferramenta, como no caso de um volume muito elevado de áudios (ex. milhões de áudios), para a verificação do comportamento da ferramenta nessas circunstâncias.

No caso da usabilidade, os testes visavam o funcionamento da ferramenta na abordagem colaborativa, sendo realizados com um pequeno número de participantes, e sua validação deve ser realizada ao longo da utilização do sistema.

5.3 TRABALHOS FUTUROS

Como trabalhos futuros, pretende-se uma maior automatização do sistema, passando a ser capaz, por exemplo, de reconhecer os ***tipos*** de ruídos/interferências e/ou situações anômalas e onde se encontram no áudio, uma vez que tais ruídos dificultam a classificação automática e isolá-los traria uma grande vantagem à classificação. Existem ainda outras questões que podem ser consideradas, como por exemplo: a importância da qualidade de gravação dos áudios, visto que, na maioria dos casos, as escolas públicas brasileiras raramente possuem equipamentos ou espaço próprios para a gravação dessas leituras. Há também que se considerar outras maneiras para a avaliação da oralidade, como por

exemplo, a interpretação e compreensão do que é lido, a capacidade de articular respostas e etc.

Adaptações dessa abordagem podem ser usadas na avaliação da aprendizagem e desenvolvimento em um novo idioma, não apenas para alunos em fase de alfabetização mas também àqueles que ainda não adquiriram a fluência.

Há também a possibilidade do desenvolvimento de modelos acústicos treinados especificamente para as faixas etárias que serão avaliadas, o que tende a melhorar o desempenho do sistema. E finalmente, a classificação de áudios que não tenham uma duração pré-estabelecida, gerando um classificador mais robusto.

REFERÊNCIAS

[1] Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan. Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–873. IEEE, 2007.

[2] Ana Marina Teixeira Araújo, Crediné Silva de Menezes, and Davidson Cury. Apoio automatizado à avaliação da aprendizagem utilizando mapas conceituais. In *Simpósio Brasileiro de Informática na Educação*, volume 1, pages 287–296, 2003.

[3] Nathalia Alves Rocha Batista et al. Estudo sobre identificação automática de sotaques regionais brasileiros baseada em modelagens estatísticas e técnicas de aprendizado de máquina. 2019.

[4] Andrew Biemiller. Relationships between oral reading rates for letters, words, and simple text in the development of reading achievement. *Reading Research Quarterly*, 1977.

[5] Matthew Black, Joseph Tepperman, Sungbok Lee, Patti Price, and Shrikanth S Narayanan. 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*, 2007.

[6] Matthew P Black, Abe Kazemzadeh, Joseph Tepperman, and Shrikanth S Narayanan. 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, 2011.

[7] Felipe Bravo-Marquez, Gaston L'Huillier, Patricio Moya, Sebastián A Ríos, and Juan D Velásquez. An automatic text comprehension classifier based on mental models and latent semantic features. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, page 23. ACM, 2011.

[8] CAEd. O que fazemos. http://institucional.caed.ufjf.br/o-que-fazemos/, May 2019.

[9] Maurício CANUTO. Leitura: um contraponto entre a fala do professor e o silenciamento da voz do aluno. *Monografia (Especialização)–Centro de Pós Graduação, Universidade Nove de Julho, São Paulo*, 2008.

[10] Luiz Carlos Carchedi. Uma abordagem tecnológica para o aprimoramento do ensino. In *Simpósio Nacional da ABCiber*, 2019.

[11] Luiz Carlos Carchedi, Eduardo Barrére, and Jairo Souza. Abordagem colaborativa para apoio à avaliação do ensino de português. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1593, 2018.

[12] Luiz Carlos Carchedi, Eduardo Soares, Jorão Gomes Jr, Eduardo Barrére, and Jairo Souza. 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, 2018.

[13] Luiz Carlos Carchedi, Jairo Souza, Eduardo Barrére, and Fabrício Mendonça. Onto4la: uma ontologia para integração de dados educacionais. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 7, page 439, 2018.

[14] Henildes José Carrer, Ednaldo Brigante Pizzolato, and Celso Goyos. Avaliação de software educativo com reconhecimento de fala em indivíduos com desenvolvimento normal e atraso de linguagem. *Brazilian Journal of Computers in Education*, 17(03):67, 2009.

[15] Mirella Ribeiro Chaer and Edite da Glória Amorim Guimarães. 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, 2012.

[16] Felix Claus, Hamurabi Gamboa Rosales, Rico Petrick, Horst-Udo Hain, and Rüdiger Hoffmann. A survey about asr for children. In *Speech and Language Technology in Education*, 2013.

[17] Catia Cucchiarini, Helmer Strik, and Lou Boves. 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, 2000.

[18] Sara da Costa Freitas, Josina Teixeira, and Miriam Machado. Desafios no ensino da oralidade. *Cadernos de Estudos e Pesquisa na Educação Básica*, 2(1):197–215, 2017.

[19] Carlos da Fonseca Brandão. *PNE passo a passo:(Lei no 10.172-2001): Discussão dos objetivos e metas do plano nacional de educação*. avercamp, 2006.

[20] Marinês Jesus da SiLva and Amanda Valiengo. O desenvolvimento da oralidade na educação infantil. *Revista Interfaces*, 2(2):21–24, 2010.

[21] MEC Lei de Diretrizes. Bases da educação nacional–lei 9.394/96. *Brasília: MEC*, 1996.

[22] Zilma de Moraes Ramos de Oliveira. *Educação Infantil: fundamentos e métodos*. Cortez Editora, 2014.

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

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

[25] Jacques Duchateau, Leen Cleuren, Hugo Van hamme, and Pol Ghesquière. Automatic assessment of children's reading level. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[26] Jerome V D'Agostino, Robert H Kelly, and Emily Rodgers. Self-corrections and the reading progress of struggling beginning readers. *Reading Psychology*, pages 1–26, 2019.

[27] Marcos Valadão Gualberto Ferreira and Jairo Francisco de Souza. Use of automatic speech recognition systems for multimedia applications. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, pages 33–36. ACM, 2017.

[28] Tais Ferreira and Marcia Fernandes. Detecção de traços de personalidade em textos para apoiar a formação de grupos para colaboração. In *Simpósio Brasileiro de Informática na Educação*, volume 28, page 1627, 2017.

[29] Raquel Meister Ko Freitag and Mônica Maria Soares Rosário. A provinha brasil na visão dos professores. *PROLÍNGUA*, 8(1), 2013.

[30] Lynn S Fuchs, Douglas Fuchs, Michelle K Hosp, and Joseph R Jenkins. Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific studies of reading*, 5(3):239–256, 2001.

[31] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. A review of asr technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, page 7. ACM, 2009.

[32] Ana Valeska Amaral Gomes. Custo aluno qualidade. *Brasília: Câmara dos Deputados*, 2009.

[33] Rainer E Gruhn, Wolfgang Minker, and Satoshi Nakamura. *Statistical pronunciation modeling for non-native speech processing*. Springer Science & Business Media, 2011.

[34] Annika Hämäläinen, Hugo Meinedo, Michael Tjalve, Thomas Pellegrini, Isabel Trancoso, and Miguel Sales Dias. Improving speech recognition through automatic selection of age group–specific acoustic models. In *International Conference on Computational Processing of the Portuguese Language*, pages 12–23. Springer, 2014.

[35] Chao Huang, Tao Chen, and Eric Chang. Accent issues in large vocabulary continuous speech recognition. *International Journal of Speech Technology*, 7(2-3):141–153, 2004.

[36] Oliver Jokisch, Horst-Udo Hain, Rico Petrick, and Rüdiger Hoffmann. Robustness optimization of a speech interface for child-directed embedded language tutoring. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, page 10. ACM, 2009.

[37] Cinthya Eiko Kawano, Adriana de Souza Batista Kida, Carolina Alves Ferreira de Carvalho, and Clara Regina Brandão de Ávila. Parâmetros de fluência e tipos de erros na leitura de escolares com indicação de dificuldades para ler e escrever. *Revista da Sociedade Brasileira de Fonoaudiologia*, 2011.

[38] Melanie R Kuhn, Paula J Schwanenflugel, and Elizabeth B Meisinger. Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, 45(2):230–251, 2010.

[39] Joseph P Magliano, Keith Millis, Yasuhiro Ozuru, and Danielle S McNamara. A multidimensional framework to evaluate reading assessment tools. *Reading comprehension strategies: Theories, interventions, and technologies*, pages 107–136, 2007.

[40] Andreas Maier, Stefanie Horndasch, and Elmar Nöth. Automatic classification of reading disorders in a single word reading test. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, page 9. ACM, 2009.

[41] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *International Conference on Spoken Language. ICSLP 96.*, volume 3, pages 1457–1460. IEEE, 1996.

[42] Sandra Puliezi and Maria Regina Maluf. A fluência e sua importância para a compreensão da leitura. *Psico-USF*, 19(3*):467–475, 2014.

[43] Martin Russell, Shona D'Arcy, and Lit Ping Wong. Recognition of read and spontaneous children's speech using two new corpora. In *Eighth International Conference on Spoken Language Processing*, 2004.

[44] Kamini Sabu, Kanhaiya Kumar, and Preeti Rao. Improving the noise robustness of prominence detection for children's oral reading assessment. In *2018 Twenty Fourth National Conference on Communications (NCC)*, pages 1–6. IEEE, 2018.

[45] Sandra Zákia Sousa and Cláudia Oliveira Pimenta. Avaliação da educação infantil: aportes de iniciativas estrangeiras. *Estudos em Avaliação Educacional*, 27(65):376–406, 2016.

[46] Linda Suskie. *Assessing student learning: A common sense guide*. John Wiley & Sons, 2018.

[47] Sheila W Valencia, Antony T Smith, Anne M Reece, Min Li, Karen K Wixson, and Heather Newman. Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, 45(3):270–291, 2010.

[48] Silza Maria Pasello Valente. A avaliação da aprendizagem no contexto da reforma educacional brasileira. *Estudos em Avaliação educacional*, pages 75–88, 2003.

[49] Katrien Verbert, Erik Duval, Joris Klerkx, Sten Govaerts, and José Luis Santos. Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10):1500–1509, 2013.

[50] Heraldo Marelim Vianna. Avaliações nacionais em larga escala: análises e propostas. *Estudos em avaliação educacional*, pages 41–76, 2003.

[51] Ramya Viswanathan, Periyasamy Paramasivam, and Jithendra Vepa. Hierarchical accent determination and application in a large scale asr system. In *Interspeech*, pages 1958–1959, 2018.

[52] Jay G Wilpon and Claus N Jacobsen. A study of speech recognition for children and the elderly. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 349–352. IEEE, 1996.

[53] Martin Wöllmer, Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Tandem decoding of children's speech for keyword detection in a child-robot interaction scenario. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(4):12, 2011.

[54] Shasha Xie, Keelan Evanini, and Klaus Zechner. 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, 2012.

[55] Yanli Zheng, Richard Sproat, Liang Gu, Izhak Shafran, Haolang Zhou, Yi Su, Daniel Jurafsky, Rebecca Starr, and Su-Youn Yoon. Accent detection and speech recognition for shanghai-accented mandarin. In *Ninth European Conference on Speech Communication and Technology*, 2005.