Avaliação Automática da Fluência em Leitura para Crianças em Fase de Alfabetização

Eduardo R. Soares[1], **Luiz Carlos Carchedi**[1], **Jorão Gomes Jr.**[1]
Eduardo Barrére[1] **e Jairo Francisco de Souza**[1]

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

{eduardosoares,carchedi,joraojunior} @ice.ufjf.br

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

Abstract. *In Brazil, large-scale learning assessment is fundamental for public bodies responsible for education. Through these evaluations, it is possible to plan public policies aimed at improving education. When it comes to the first years of elementary school, an important aspect to evaluate is the ability of the students to use their mother tongue in the oral mode. Nevertheless, the process of evaluating orality on a large scale is still a very costly and time-consuming task. This paper proposes and evaluates the use of Automatic Speech Recognition (ASR) for the automation of this evaluation process. Experiments were performed on a real base of audios and it was demonstrated that the automatic evaluation closely reflects the quality of the analyzed readings.*

Resumo. *No Brasil, a avaliação de aprendizagem em larga escala é fundamental para os órgãos públicos responsáveis pela educação. Quando se trata dos primeiros anos do ensino fundamental, um importante aspecto a se avaliar é a capacidade do aluno utilizar a sua língua materna na modalidade oral. Apesar disso, o processo de avaliação da fluência em leitura em larga escala ainda é uma tarefa muito dispendiosa em termos financeiros e de tempo. Este trabalho propõe e avalia o uso de reconhecimento automático de fala (ASR) para a automatização desse processo. Experimentos foram realizados em uma base real de áudios e foi demonstrado que a avaliação automática reflete, de forma próxima, a qualidade das leituras analisadas por especialistas.*

1. Introdução

A avaliação de aprendizagem é um processo fundamental para o ensino, pois através dela é possível corroborar de forma significativa a qualidade do mesmo. Com base nessas avaliações, os educadores são capazes de analisar o nível de compreensão de um conteúdo por parte de seus alunos e, dessa forma, conhecer os pontos deficitários de uma turma para a definição de metas e planos de ensino [Suskie 2018].

Assim como a avaliação de aprendizagem faz parte do processo educacional do dia-a-dia da sala de aula, geralmente órgãos institucionais, como o Ministério da Educação (MEC), costumam empregar avaliações em larga escala para todos os níveis de

DOI: 10.5753/cbie.sbie.2018.11

ensino (fundamental, médio, superior) como forma de obter uma visão macroscópica da qualidade da educação no país e, dessa forma, planejar políticas públicas para suprir as deficiências detectadas [Sousa and Arcas 2010]. Geralmente, esse tipo de avaliação ocorre através de provas escritas que são aplicadas nacionalmente a fim de avaliar a competência dos alunos nos assuntos que os próprios órgãos de ensino definem. Apesar disso, algumas competências definidas não são passíveis de serem avaliadas através de provas escritas e acabam ficando de fora dessas avaliações, como é o caso da fluência em leitura. Mesmo que o próprio MEC inclua a avaliação da fluência em leitura para o ensino fundamental em seus Parâmetros Nacionais Curriculares e na Base Nacional Comum Curricular, a modalidade oral da língua costuma ser preterida em relação à escrita nas escolas, assim como nas avaliações em larga escala que não a contemplam [da Costa Freitas et al. 2017]. Muito disso deve-se ao fato de que avaliar esse tipo de competência é um processo custoso em termos financeiros e de tempo. Portanto, é interessante que esse processo possa ser automatizado a fim de viabilizar sua aplicação em larga escala.

Sendo assim, devido à importância da fluência em leitura nos primeiros anos do ensino fundamental e à falta de mecanismos eficientes para sua avaliação em larga escala, o presente trabalho tem como objetivo discutir aspectos do uso de técnicas de reconhecimento automático de fala (ASR) para a avaliação da fluência em leitura, que é um dos principais aspectos da oralidade, em crianças na fase de alfabetização. Assim, neste trabalho, mostra-se que é possível obter automaticamente métricas que reflitam, de forma muito satisfatória, a qualidade de leituras gravadas em formato digital dessas crianças.

Visto que os trabalhos da literatura, até então, propõem apenas o uso de ASR para a avaliação de fluência para falantes de uma segunda língua, este trabalho inova ao propor e investigar a utilização de ASR para a avaliação automática de fluência em leitura para crianças que estão na fase de alfabetização de sua língua materna. Nesse cenário, vários desafios se fazem presentes e precisam ser considerados. Além disso, a automatização tende a tornar o processo de avaliação mais ágil e com taxas de erro aceitáveis, podendo ser uma forma eficaz de realizar esse tipo de avaliação em larga escala. Como forma de demonstrar a aplicabilidade de ASR para essa tarefa, são apresentados os resultados de experimentos realizados em uma base real de áudios. Além disso, durante este trabalho, são discutidos os desafios e especificidades ao utilizar tal abordagem direcionada ao público em questão.

Este trabalho está organizado da seguinte forma. Na Seção 2 é apresentada uma revisão da literatura que aborda aspectos relevantes em relação ao uso de ASR para o público infantil e trabalhos relacionados. A Seção 3 detalha o desenvolvimento da abordagem utilizada neste trabalho. Na Seção 4 são apresentados os experimentos e seus resultados. Finalmente, na Seção 5 as conclusões e trabalhos futuros são apresentados.

2. Revisão Bibliográfica

Nesta seção, é feita uma discussão acerca das particularidades presentes no uso de técnicas de ASR para crianças. Por fim, são apresentados os trabalhos existentes na literatura que corroboram a relevância do tema abordado neste trabalho.

2.1. Reconhecimento automático de fala infantil

O ASR é um conjunto de técnicas que tem como objetivo a conversão em texto da fala contida em um sinal de áudio [Gruhn et al. 2011]. Na abordagem mais utilizada, são infe-

ridas palavras através de uma análise do sinal de áudio e por meio do cálculo da probabilidade de uma dada palavra estar associada a um trecho de áudio de acordo com o seu som e sua frequência de ocorrência na língua [Gruhn et al. 2011]. Em geral, na construção de um sistema de ASR utiliza-se um decodificador para a interpretação do sinal de áudio e identificação; e também de modelagens acústicas, léxicas e de linguagem. O modelo acústico permite que o sistema associe partes do sinal de voz à unidades de fala (geralmente fonemas). O léxico é responsável pelo mapeamento de sequências de fonemas em palavras de um dicionário. Por fim, o modelo de linguagem é responsável por impedir que frases agramaticais sejam formadas, atribuindo probabilidades para ocorrência de sequências de palavras. Dessa forma, é possível mapear grande parte da estrutura linguística de um idioma específico para que o sistema consiga reconhecer o que foi dito.

Existem trabalhos na literatura que investigam o uso de ASR para voz infantil. É um consenso entre eles que a acurácia do reconhecimento de fala infantil geralmente é mais baixa que em adultos [Claus et al. 2013]. Essa baixa acurácia se deve às características específicas existentes na fala infantil que as diferem dos adultos. 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. Outra diferença comumente apresentada na literatura diz respeito às habilidades linguísticas infantis, principalmente em relação à dificuldade de pronúncia de certos fonemas [Claus et al. 2013].

À medida que as investigações acerca de ASR em crianças foram avançando, também foram surgindo formas de lidar com os desafios envolvidos nessa tarefa. Em [Wilpon and Jacobsen 1996], 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 [Russell et al. 2004], 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. A grande desvantagem desse método é a escassez de bases de treinamento, o que dificulta a criação de um modelo bem treinado para cada faixa de idade, fazendo com que apenas um modelo genérico sem especificidade de faixa etária seja criado [Claus et al. 2013]. Por outro lado, existem trabalhos na literatura que demonstraram que 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 [Jokisch et al. 2009].

Ainda, em [Claus et al. 2013] é relatado que a modelagem de pronúncia pode aumentar a acurácia desses reconhecedores de fala infantil quando o mesmo tem seu modelo acústico treinado com adultos. Isso se deve ao fato de que crianças geralmente possuem uma pronúncia ainda pobre, muitas vezes não utilizando a pronúncia padrão das palavras. Por isso, uma forma de melhorar a acurácia desses reconhecedores é através da adição das formas alternativas de pronúncia no modelo léxico.

2.2. Trabalhos Relacionados

Existem várias abordagens na literatura que utilizam o ASR como uma ferramenta para avaliação automática de leitura e fala. Muito dessas 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 [Neumeyer et al. 1996, Cucchiarini et al. 2000, Xie et al. 2012]. Em [Neumeyer et al. 1996], é apresentado um sistema para a avaliação da pronúncia da língua francesa por parte de 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, obtém a pontuação de forma automática. Para a realização dos experimentos, professores franceses atribuíram notas para a pronúncia de alunos não-nativos e essas notas foram comparadas às 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 profissionais. Isso mostra que a hipótese levantada pelos autores de que alunos mais avançados no estudo da língua tendem a falar mais rápido que iniciantes pode ser um bom indicador de fluência.

Já em [Cucchiarini et al. 2000] é apresentada uma avaliação quantitativa da fluência de estudantes de uma segunda língua através de tecnologias de ASR. Para isso, foi conduzido um experimento similar ao visto em [Neumeyer et al. 1996], onde as avaliações de especialistas foram comparadas às atribuídas automaticamente pelo sistema. A principal diferença entre os dois trabalhos é que [Cucchiarini et al. 2000] utilizou uma gama maior de métricas. Apesar disso, ambos concluíram que o número de palavras por período de tempo foi a métrica que mais refletiu a avaliação dada pelos especialistas.

Em [Xie et al. 2012], diferente dos trabalhos de [Neumeyer et al. 1996] e [Cucchiarini et al. 2000], é apresentada uma abordagem de ASR para classificar automaticamente a proficiência em fala espontânea em uma língua estrangeira levando em consideração informações de mais alto nível. Essas informações consistem no conteúdo da fala, relevância na atualidade e estrutura e informação do discurso. A hipótese dos autores na realização desse trabalho foi a de que boas redações se assemelham umas às outras na escolha de palavras. Assim, isso também deveria se aplicar às respostas orais. Na proposta dos autores, 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 havia erros nas transcrições geradas pelo ASR, caso o contrário, isso não era observado.

O uso de técnicas de ASR também já foi explorado na área clínica. Por exemplo, em [Maier et al. 2009] e [Carrer et al. 2009] é investigada a utilização de ASR para avaliação automática da fala e leitura em 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 ainda faltar uma avaliação mais extensiva dos métodos.

Como apresentado, inúmeros desafios estão associados ao ASR, principalmente quando aplicado para avaliar a capacidade de leitura. Embora o uso de ASR para avaliação de leitura tem se dado, principalmente, para avaliação do aprendizado de segunda língua,

estes métodos podem ser utilizados também para avaliação automática da fluência em leitura de crianças em fase de alfabetização de sua primeira língua. Neste sentido, este trabalho discute os aspectos tecnológicos presentes neste tipo de avaliação e apresenta dados de experimentos que foram realizados em um *corpus* real de áudios a fim de demonstrar o seu comportamento e potencial de ser empregado em avaliações em larga escala.

3. Desenvolvimento

Este trabalho consiste na utilização de ASR aplicado a áudios contendo leituras de crianças em fase de alfabetização para a obtenção automática de métricas de fluência que possam classificar de forma satisfatória a capacidade de leitura desses indivíduos. Com isso, deseja-se mostrar que tecnologias como essa podem ser empregadas como ferramentas de auxílio neste tipo de avaliação, fazendo com que o processo de avaliar a fluência em leitura da primeira língua seja uma tarefa menos custosa. O sistema de ASR utilizado foi o descrito em [Ferreira and de Souza 2017].

No cenário de aplicação deste trabalho, as crianças são instruídas pelo professor a ler um texto-base, enquanto a leitura é gravada por um telefone celular e posteriormente enviada para processamento. Vale ressaltar que este cenário de aplicação é bastante desafiador para sistemas de ASR. Além das questões levantadas na Seção 2, outros fatores tais como a baixa qualidade de gravação do áudio, fala do professor e ruídos de fundo que vazam no áudio, podem tornar a acurácia do ASR ainda menor. Para contornar esses problemas, optou-se pela utilização de um algoritmo de alinhamento temporal forçado [McAuliffe et al. 2017] em conjunto com o ASR. O alinhamento forçado consiste em alinhar temporalmente o sinal de áudio com o texto base de leitura, ou seja, determinar o tempo de ocorrência e duração de cada palavra e fonema do texto-base que foi lido no áudio. Para o reconhecimento das sequências de fonemas, o alinhador utiliza o modelo acústico treinado do próprio sistema ASR. Já para o mapeamento das sequências de fonemas em palavras, é utilizado um dicionário léxico próprio, que pode ser canônico ou um dicionário alterado, a fim de reconhecer variações fonéticas na fala. Por fim, o alinhador constrói um modelo de linguagem utilizando o texto de leitura como referência. Dessa forma, limitam-se as probabilidades de reconhecimento de palavras àquelas que ocorrem no texto, forçando o alinhamento do áudio com o texto e, assim, diminuindo a chance de que os fatores acima interfiram no reconhecimento.

O resultado do reconhecimento de fala via alinhamento temporal forçado é utilizado para gerar métricas quantitativas de fluência que, em seguida, são usadas para classificar a leitura contida nos áudios. É importante ressaltar que o uso do dicionário de pronúncias padrão no alinhador forçado pode fazer com que o sistema force o alinhamento de palavras pronunciadas de forma diferente do padrão aceitável na língua. Essa anomalia pode ocorrer quando as sonoridades das duas pronúncias (incorreta e correta) são parecidas. Na maior parte dos casos, esse fenômeno não afeta de maneira significativa as métricas de fluência geradas pelo sistema. Porém, caso haja a necessidade de identificar esses casos, uma forma de fazê-lo é alterar o modelo léxico utilizado pelo alinhador de forma a inserir pronúncias alternativas para determinadas palavras. Dessa forma, quando alguma dessas pronúncias alternativas é detectada no áudio, é possível identificar qual erro de pronúncia ocorreu. Apesar disso, a alteração do léxico costuma ser um processo dispendioso que precisa ser feito manualmente, se tornando inviável em alguns casos.

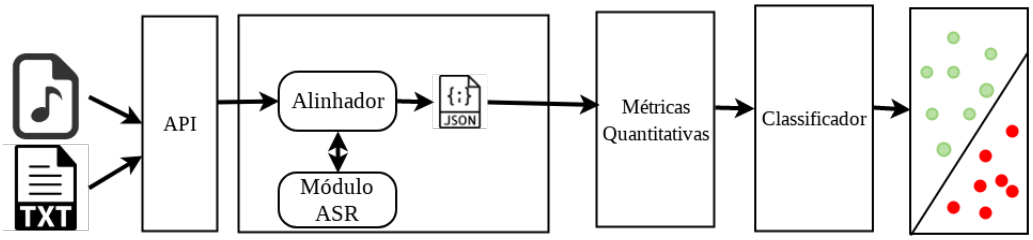

Figura 1. Arquitetura do sistema de avaliação automática de oralidade

Para avaliação deste trabalho, foi necessário o desenvolvimento de uma arquitetura de *software* que consiste de uma interface de entrada para receber o texto-base e um arquivo de áudio contendo a leitura desse texto. Além disso, foi construído um módulo ASR junto com um alinhador forçado no núcleo do sistema. Como saída do processamento do áudio junto com o texto de leitura, são geradas métricas quantitativas de fluência associadas ao áudio de entrada. Essa métricas são utilizadas por um classificador para determinar se a leitura foi satisfatória ou não. O esquema lógico de funcionamento do sistema desenvolvido pode ser visto na Figura 1.

3.1. Métricas quantitativas

As métricas quantitativas de fluência, obtidas de forma automática, são o objetivo principal do sistema. Através delas é possível obter uma avaliação objetiva e que represente de forma bem próxima a qualidade de leituras de crianças que estão nos primeiros anos do ensino fundamental. O sistema implementado é capaz de gerar duas métricas: quantidade de palavras lidas (QPL) e taxa de palavras lidas por tempo (TPL).

3.1.1. Quantidade de palavras lidas (QPL)

A QPL mede a quantidade de palavras do texto que foram lidas corretamente pelo leitor e é calculada como o número de palavras do texto base que foram reconhecidas pelo sistema na ordem correta. Nessa métrica, assume-se que se o sistema conseguiu reconhecer a pronúncia do leitor para uma palavra w_i, então o leitor foi capaz de pronunciá-la com o mínimo de correção esperada, de acordo com o dicionário léxico (de pronúncias) utilizado. Por outro lado, se o leitor pronunciar a palavra com sonoridade muito distante do esperado, o sistema não será capaz de reconhecê-la, ou seja, o valor dessa métrica tenderá a cair para casos de pronúncias muito ruins. O cálculo do QPL pode ser visto na Equação 1, onde w_i é a palavra do texto na posição i e r_i é a palavra reconhecida na mesma posição. Γ denota o arranjo de palavras que representa o texto e Υ o arranjo de palavras reconhecidas pelo sistema. O valor de $C(w_i)$ é 1 se $w_i = r_i$ e 0 caso contrário.

$$QPL = \sum_{1}^{N} C(w_i),\ w_i \in \Gamma \tag{1}$$

3.1.2. Taxa de palavras lidas por período de tempo (TPL ou *Rate of Speech)*

Medida utilizada para mensurar a velocidade com a qual determinado leitor consegue pronunciar corretamente as palavras contidas em um texto específico. Essa métrica já

foi amplamente explorada nos trabalhos da literatura, como visto na Seção 2, e demonstrou ser uma das mais eficientes para avaliação automática de fluência dos leitores. Essa métrica é aplicável para este trabalho, pois leva em consideração não só a capacidade da criança pronunciar corretamente as palavras, mas também a duração da leitura. Indivíduos dos primeiros anos do ensino fundamental com dificuldades na leitura, frequentemente, têm a fala marcada pela ocorrência de muitas pausas. Assim, esta métrica é um bom classificador, pois um bom leitor tende a ter, além de uma pronúncia foneticamente mais correta, um ritmo de leitura mais adequado, o que é refletido positivamente nesta métrica. O cálculo da mesma se dá através da divisão do número de palavras lidas da Equação 1 pela duração da leitura.

4. Experimentos e Resultados

A fim de evidenciar a aplicabilidade de ASR na avaliação automática da oralidade de crianças nos anos inciais de ensino, foram realizados dois experimentos. O primeiro é um estudo de caso realizado com o intuito de verificar a capacidade do sistema de identificar erros de pronúncia comumente cometidos em leituras infantis. Já no segundo experimento, as métricas de fluência geradas automaticamente são utilizadas para classificar leituras infantis reais. Essas classificações foram comparadas às feitas perceptivamente por humanos, como forma de avaliar o desempenho da abordagem.

4.1. Reconhecimento de erros de pronúncia

Neste experimento, o objetivo é verificar a capacidade do sistema reconhecer erros de leitura comumente cometidos por crianças em fase de alfabetização. Para isso, nos baseamos nas classes de erros definidas por [Kawano et al. 2011], onde identificamos 9 classes de erros passíveis de resolução pelo ASR dentre as 10 definidas pelo autor. São elas: **T1** - Troca por palavra visualmente similar (leitura de palavras aparentemente iguais); **T2** - Regularizações (palavras erradas lidas como certas); **T3** - Desrespeito à regra de correspondência grafo-fonêmica independente do contexto (substituição de consoantes ou vogais durante a pronúncia das palavras); **T4** - Omissões e adições (leituras adicionais ou esquecimento de vogais ou consoantes); **T5** - Falhas de aplicação de regras ortográficas (quando não há respeito as regras de ortográficas); **T6** - Inversões de sequências (leitura de palavras desrespeitando a ordem sequencial das letras); **T7** - Erro quanto ao emprego da tonicidade (erro na identificação da sílaba tônica); **T8** - Erro por desrespeito ao sinal gráfico de acentuação (quando a acentuação de uma palavra é ignorada); **T9** - Erros complexos (existência de mais de um erro para uma mesma palavra). A classe **T10** - Recusas (quando a criança se recusa a ler uma das palavras apresentadas no texto), por ser dependente da opção de cada indivíduo, não foi considerada para simulação do sistema.

Para validar a aplicabilidade da abordagem no reconhecimento de tais erros, foram gravados áudios contendo a leitura de um texto base também definido por [Kawano et al. 2011]. Erros das classes selecionadas foram propositalmente inseridos nos áudios. Além disso, para que o sistema conseguisse reconhecer a ocorrência desses erros, foi necessário realizar alterações no modelo léxico do sistema. Essas alterações consistem em inserir, para cada palavra do léxico que está presente no texto, pronúncias alternativas que representam os erros fonéticos que deseja-se reconhecer no áudio. Dessa forma, ao aplicar o ASR nos áudios gravados, o sistema foi capaz de reconhecer e indicar todos os erros de leitura presentes nos áudios. Portanto, com esse experimento foi

possível evidenciar que erros comuns de leitura em crianças podem ser capturados com sucesso pelo sistema mediante adaptações feitas no modelo léxico. Essas adaptações podem ser facilmente inseridas quando se conhece o texto base de avaliação e os erros que se deseja capturar, o que é viável de ser feito em avaliações onde o texto é conhecido.

4.2. Avaliação automática de fluência em uma base real

Neste experimento, a abordagem foi aplicada a uma base real de áudios composta por 160 arquivos com boa qualidade de gravação contendo leituras de crianças em fase de alfabetização. Os áudios foram captados por professores de escolas públicas através de um telefone celular e gravados em formato "***.mp3***". Ao todo, 10 textos foram lidos por essas crianças. Porém, para a realização deste experimento, nenhum tipo de separação dos áudios por texto lido, idade, gênero ou região do país foi realizado.

Em seguida, cada um desses áudios passou por uma avaliação perceptual humana. Nessa avaliação, cada um dos 3 avaliadores ficou responsável por ouvir uma parte do conjunto de áudios e classificar a leitura como **boa** ou **ruim**, de acordo com sua percepção sobre a mesma. Para a realização dessa avaliação perceptual, alguns critérios foram adotados: (1) Leituras onde eram percebidos apenas pequenos erros de pronúncia e que apresentavam uma taxa de articulação de palavras perceptivelmente boa deveriam ser classificadas como boas; (2) Leituras onde eram percebidos muitos erros de pronúncia e uma taxa de articulação aquém do esperado foram classificadas como ruins. Nessa classificação manual, 85 leituras foram classificadas como **boas** e 75 como **ruins**. Após a classificação perceptual humana, todos esses mesmos áudios foram submetidos ao processamento do sistema desenvolvido para a obtenção automática das métricas quantitativas descritas na Seção 3. Uma situação de contorno encontrada neste experimento é que todos os áudios fornecidos para a realização do mesmo tinham suas durações limitadas a 1 minuto. Portanto, como não há variação no tempo dos áudios, a métrica TPL acaba sendo igual a QPL, neste caso. Dessa forma, neste experimento as duas métricas serão tratadas como apenas TPL daqui pra frente. Assim sendo, depois de obtida a métrica TPL para a leitura de cada áudio, a mesma é dada como entrada para um classificador que determina se aquela é uma leitura **boa** ou **ruim**. Para fins de experimentos, foi gerada uma classificação binária, de forma automática, através de um *threshold*, de acordo com a seguinte fórmula:

$$Leitura_y = \begin{cases} \text{boa}, \ se \ \ TPL_y \geq k \\ \text{ruim}, \ se \ \ TPL_y < k \end{cases}$$

Para encontrar o melhor valor de k para a classificação das leituras dessa base, variou-se o valor de k entre 0 e 170, e as classificações obtidas automaticamente na base para cada valor de k foram comparadas com as feitas perceptualmente por avaliadores humanos através do cálculo do coeficiente de correlação de Pearson [Xie et al. 2012]. A Figura 2 mostra o gráfico da correlação obtida entre a avaliação humana e a automática para cada valor de k. Como pode-se notar, o melhor valor de k para a classificação dessa base foi de 73 palavras lidas por minuto. Com esse valor de k, obteve-se uma correlação de 81,20% entre a avaliação automática e perceptual. Esse alto valor de correlação obtido é uma forte evidência de que aspectos da oralidade de crianças em fase de alfabetização, como a fluência em leitura, podem ser automaticamente avaliados via ASR, o que é o principal objetivo deste trabalho. Esse valor de correlação indica que a avaliação automática se aproximou de forma bastante satisfatória da avaliação feita por humanos.

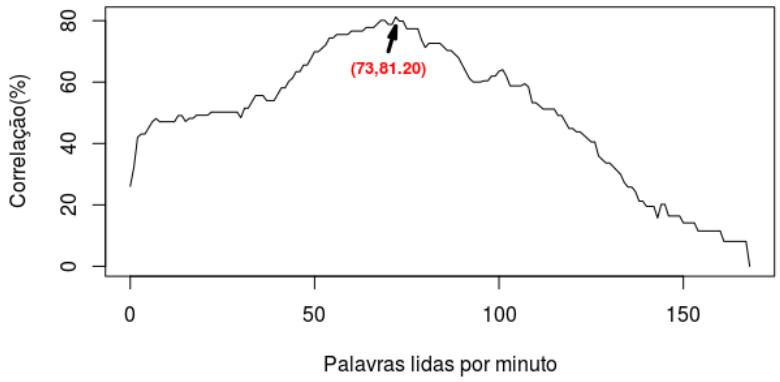

Figura 2. Correlação entre a avaliação humana e a gerada automaticamente

5. Conclusões e Trabalhos Futuros

Neste trabalho foram discutidos os principais aspectos tecnológicos que devem ser levados em consideração ao utilizar ASR para a avaliação automática da fluência de leitura de crianças em fase de alfabetização. Através dos experimentos, pode-se evidenciar o grande potencial da utilização de tal abordagem para a automatização desse processo em um cenário real de aplicação.

O uso de uma solução computacional pode tornar o processo mais rápido, propiciando uma maior quantidade de análises por período de tempo e, assim, diminuir os custos do processo de avaliação, pois com um servidor de baixo custo é possível realizar a tarefa de diversos avaliadores. Desta forma, a intervenção dos avaliadores humanos ocorrerá somente no caso de uma análise por amostragem para posterior comparação e a análise dos casos de impossibilidade na definição automática da avaliação, ou seja, para os áudios que estejam na região muito próxima a k, e que necessitam de uma avaliação humana. Todo esse processo pode ser feito de forma bem mais ágil em relação à dependência de disponibilidade somente de avaliadores humanos.

Como trabalhos futuros, propõe-se a análise de um modelo acústico treinado especificamente para crianças afeta a avaliação automática. Posteriormente, pretende-se aplicar as métricas geradas para classificar áudios que não tenham uma duração pré-estabelecida, gerando um classificador mais acurado utilizando as duas métricas geradas pelo sistema. Além disso, outras questões que ficaram de fora deste trabalho serão posteriormente discutidas, por exemplo: o peso da qualidade de gravação dos áudios sobre todo o processo, visto que, na maioria dos casos, as escolas públicas brasileiras raramente possuem equipamentos ou espaço próprios para a gravação dessas leituras.

6. Agradecimentos

Agradecemos ao CAEd/UFJF por fornecer a base de avaliação usada nos experimentos. Agradecemos também à CAPES e ao PGCC/UFJF pelo apoio financeiro.

Referências

Carrer, H. J., Pizzolato, E. B., and Goyos, C. (2009). 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.

Claus, F., Gamboa Rosales, H., Petrick, R., Hain, H.-U., and Hoffmann, R. (2013). A survey about asr for children. In *Speech and Language Technology in Education*.

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

da Costa Freitas, S., Teixeira, J., and Machado, M. (2017). Desafios no ensino da oralidade. *Cadernos de Estudos e Pesquisa na Educação Básica*, 2(1):197–215.

Ferreira, M. V. G. and de Souza, J. F. (2017). 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.

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

Jokisch, O., Hain, H.-U., Petrick, R., and Hoffmann, R. (2009). 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.

Kawano, C. E., Kida, A. d. S. B., Carvalho, C. A. F. d., and Ávila, C. R. B. d. (2011). 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*.

Maier, A., Horndasch, S., and Nöth, E. (2009). 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.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: trainable text-speech alignment using kaldi. In *Proceedings of interspeech*.

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

Russell, M., D'Arcy, S., and Wong, L. P. (2004). Recognition of read and spontaneous children's speech using two new corpora. In *Eighth International Conference on Spoken Language Processing*.

Sousa, S. Z. and Arcas, P. H. (2010). Implicações da avaliação em larga escala no currículo: revelações de escolas estaduais de são paulo. *Educação: Teoria e prática*, 20(35):181.

Suskie, L. (2018). *Assessing student learning: A common sense guide*. John Wiley & Sons.

Wilpon, J. G. and Jacobsen, C. N. (1996). 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.

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