Uso de Alinhadores Forçados para Avaliação Automática em Larga Escala da Fluência em Leitura

Jorão Gomes Jr.[1], **Warley Almeida Silva**[2], **João Victor de Souza**[2], **Eduardo Barrére**[1], **Jairo Francisco de Souza**[1]

[1] Programa de Pós-Graduação em Ciência da Computação – (UFJF)
36.360-900 – Juiz de Fora – MG – Brasil

[2]Bacharelado em Ciência da Computação – Universidade Federal de Juiz de Fora (UFJF)
36.360-900 – Juiz de Fora – MG – Brasil

{joraojunior,warley.almeida,joao.souza}@ice.ufjf.br

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

Abstract. *The use of automatic speech recognition (ASR) systems has grown significantly in recent years. However, there are not many works that use ASR techniques for evalute reading fluency in literacy. The use of transcription as the only means of evaluation brings some problems by masking the incorrect pronunciation of some words. Therefore, this work proposes the use of forced alignment algorithms combined with the ASR system. The results show that our approach is faster and exceeds the accuracy of others by 22,11% in a real dataset containing orality tests.*

Resumo. *A utilização de sistemas para reconhecimento automático de fala (ASR) tem crescido bastante nos últimos anos. No entanto, não existem muitos trabalhos que utilizam técnicas de ASR para a avaliação de fluência em leitura na alfabetização. A utilização da transcrição como único meio de avaliação traz alguns problemas por mascarar a pronúncia incorreta de algumas palavras. Assim, este trabalho propõe o uso de algoritmos de alinhamento forçado em conjunto com o sistema de ASR. Os resultados mostram que a abordagem proposta alcança resultados mais rápidos e com acurácia 22,11% maior em uma base real de áudios provenientes de avaliações de oralidade, mostrando-se a alternativa mais aplicável para a avaliação em larga escala.*

1. Introdução

O número de aplicações que fazem uso de sistemas de reconhecimento automático de fala (ASR – *Automatic Speech Recognition*) presentes no cotidiano tem crescido bastante nos últimos anos, como pode ser visto na popularidade dos assistentes pessoais virtuais dos *smartphones* [Hoy 2018]. Isso se deve ao desenvolvimento das tecnologias envolvidas na interpretação dos sinais sonoros e na melhoria da acurácia desses sistemas. As técnicas de ASR permitem obter inúmeras informações provenientes da fala. Entre elas, quais palavras foram pronunciadas correta e incorretamente pelo falante, transcrições fonéticas, etc [Soares et al. 2018]. Por esse motivo, técnicas de ASR já foram exploradas e implementadas com sucesso para a avaliação da fluência em leitura de indivíduos no aprendizado de línguas estrangeiras [Demenko et al. 2010, Thomson 2011]. No entanto, embora não

DOI: 10.5753/cbie.sbie.2019.61

existam muitos trabalhos que utilizam técnicas de ASR para a avaliação de fluência em leitura na alfabetização, é possível a utilização do conjunto de informações extraídas de um áudio para atingir esse objetivo.

Segundo [Fuchs et al. 2001], a fluência em leitura pode ser compreendida como a habilidade com que um individuo traduz oralmente um texto com velocidade e precisão através da rapidez no reconhecimento das palavras e a segmentação correta dos fonemas pronunciados. Nesse sentido, o Ministério da Educação (MEC), através da Base Nacional Comum Curricular (BNCC), determina que uma capacidade mínima de correta decodificação das palavras seja adquirida para cada faixa etária [MEC 2018]. Visto que a fluência em leitura está estritamente ligada à habilidade de se pronunciar, há a necessidade de se monitorar e avaliar de forma adequada a maneira com que a língua oral está sendo desenvolvida.

É relevante a necessidade de se avaliar a fluência, uma vez que nem sempre é possível suprir as dificuldades de todos os alunos dentro da sala de aula, deixando defasado o desenvolvimento da linguagem oral [Celestino 2019]. Ainda, ao se trabalhar com avaliações em escala nacional, são necessárias as definições de padrões de correção, armazenamento e administração de grandes volumes de dados, admissão de profissionais capacitados e produção de resultados em tempo adequado [Carchedi et al. 2018]. Para crianças na fase de alfabetização, uma avaliação automática da fluência torna-se ainda mais desafiadora por ter que lidar com crianças em faixas etárias distintas ainda no desenvolvimento da alfabetização, emergindo as dificuldades de pronúncia de alguns fonemas e variação do timbre da voz, por exemplo.

Mesmo com todos os desafios no processo de avaliação de fluência, a aplicação de técnicas de ASR se mostra promissora. É possível adaptar sistemas de reconhecimento de fala para forçar o reconhecimento de erros de pronúncia através da utilização de alinhadores forçados. Sendo pequena a exploração de técnicas que utilizem sistemas de ASR para avaliação automática da fluência em língua nativa, principalmente para crianças na fase de alfabetização, este trabalho propõe uma abordagem para a automatização desse processo utilizando sistemas ASR em conjunto com algoritmos de alinhamento temporal forçados [McAuliffe et al. 2017]. O trabalho está estruturado da seguinte forma: a Seção 2 apresenta a revisão bibliográfica sobre o tema, a Seção 3 descreve a proposta utilizada, a Seção 4 apresenta os experimentos realizados e os resultados alcançados e, por fim, a Seção 5 apresenta as conclusões e os trabalhos futuros.

2. Revisão Bibliográfica

Nesta seção são apresentados os principais temas envolvendo a utilização de ASR e avaliação da fluência em leitura. Em seguida, são levantados os trabalhos relacionados com a abordagem proposta.

2.1. Sistemas ASR e avaliação de fluência em leitura

Por definição, sistemas de ASR têm por finalidade converter sinais de áudio em texto. Para alcançar este objetivo, tais sistemas utilizam uma arquitetura composta por um sinal de entrada, um decodificador e modelos acústico, léxico e de linguagem [Yu and Deng 2016, Ferreira and de Souza 2017]. O **sinal de entrada** é o áudio que será processado. O **decodificador** é responsável por analisar diretamente o sinal de áudio de

entrada e gerar a transcrição do áudio com o auxílio dos três modelos mencionados. O **modelo acústico** é responsável pela atribuição de uma unidade da linguagem, geralmente um fonema, a um trecho de sinal sonoro. Dessa forma, o modelo acústico interpreta computacionalmente o áudio como uma onda e para cada trecho de onda associa um fonema específico conforme suas características. Após a conversão de todos os trechos de onda em fonemas, a sequência de fonemas obtida é convertida em palavras, de acordo com as palavras reconhecidas pelo modelo léxico. O **modelo léxico** consiste em um dicionário onde cada entrada é uma palavra acompanhada da sua respectiva transcrição fonética. Por fim, o **modelo de linguagem** é responsável por garantir que as frases transcritas façam sentido gramaticalmente, ou seja, que frases agramaticais não sejam formadas consistindo num conjunto de probabilidades de que pares de palavras estejam seguidas uma da outra no modelo. Com base nessas probabilidades, é possível decidir entre possíveis transcrições para um conjunto de fonemas e gerar transcrições que façam sentido.

Esses tipos de sistemas podem ser adaptados e utilizados para avaliação de fluência em larga escala. Segundo [Hudson et al. 2005], para se obter a fluência em leitura existem três elementos principais que uma pessoa deve possuir: uma precisão na leitura, boa taxa de leitura de palavras e possuir prosódia durante a pronúncia. Dessa forma, um leitor fluente deve ser capaz de manter, por uma certa quantidade de tempo, uma leitura precisa e um conhecimento correto sobre as palavras faladas. Embora a avaliação de fluência em larga escala auxilie na identificação de problemas na educação e na tomada de decisões pelos gestores, o custo da sua realização não pode ser desconsiderado. A principal vantagem no uso de sistemas ASR neste tipo de avaliação é a diminuição de custos operacionais para geração dos resultados finais, visto que reduz-se drasticamente a quantidade de corretores humanos necessários para o processo. Por outro lado, o uso de sistemas computacionais como este não são isentos de erros e precisam ser cuidadosamente construídos para garantir uma taxa de acerto dentro da faixa de tolerância definida. É possível adaptar sistemas ASR para monitorar e avaliar a língua nativa da seguinte forma: (i) um modelo acústico pode ser treinado com gravações especializadas para o cenário de aplicação; (ii) um modelo de linguagem pode ser construído com base nas estruturas frasais dos textos que serão avaliados e (iii) um dicionário léxico pode ser definido para validação das pronúncias realizadas.

Das informações possíveis de serem geradas por esse tipo de análise do áudio, três são importantes para a avaliação de fluência: quantidade de palavras lidas (QPL), quantidade de palavras lidas corretamente (QPC) e precisão de leitura. Considerando um texto conhecido, o (QPL) é definido como a quantidade de palavras pronunciadas pela criança ao longo da sua leitura. O QPC, por sua vez, representa apenas a quantidade de palavras lidas que estão efetivamente no texto de referência e de acordo com a pronúncia esperada. Por fim, a precisão da leitura é calculada como a razão do QPC pelo QPL.

2.2. Trabalhos Relacionados

Sistemas para ASR têm sido largamente utilizados para classificação de pronúncia de segunda língua [Neri et al. 2003, Neri et al. 2006, Campos and Freitas 2016, Litman et al. 2018]. Em geral, os objetivos dessas iniciativas são mensurar a qualidade como uma língua estrangeira está sendo desenvolvida e administrar os principais erros cometidos pelos falantes em fase de desenvolvimento. Em [Neri et al. 2006], por exemplo, é construído um sistema ASR para prover *feedback* das pronúncias

realizadas em Holandês. A abordagem tem como entrada uma pronúncia realizada e o *feedback* consiste na identificação dos fonemas pronunciados erroneamente. Já em [Campos and Freitas 2016], o foco se incrementa tendo como alvo não só a identificação de erros de pronúncias, mas também a utilização pedagógica em instituições de ensino superior de sistemas ASR para a aprendizado da língua inglesa como segunda língua.

No cenário de avaliação da leitura de crianças, os desafios para adaptação de sistemas ASR se intensificam. [Claus et al. 2013] mostra que a acurácia de sistemas de ASR é geralmente menor para vozes infantis do que para vozes adultas. Isso pode ser explicado pela dificuldade em pronunciar certos fonemas, uma vez que a linguagem ainda está sendo desenvolvida além das diferenças entre o trato vocálico de crianças e adultos, de forma que o timbre e outros aspectos da voz variam com frequência. Além disso, outro grande problema é falta de bases para criação de modelos acústicos específicos para lidar com aplicação específicas para crianças. Por conta disso, trabalhos como [Liao et al. 2015, Yeung and Alwan 2018] tentam contornar essas dificuldades através do treinamento de redes neurais utilizando adaptações de bases de áudios construídas com falas de adultos para identificação de pronúncias realizadas por crianças.

Mesmo com essas limitações, é possível afirmar que a aplicação de técnicas de ASR para avaliação automática de fluência em leitura de crianças tem avançado em qualidade. Para contornar as limitações das bases de áudios por crianças, é possível utilizar algoritmos que forçam a identificação temporal das pronúncias realizadas e as alinham com textos de referência [Eskenazi 1996, Moreno et al. 1998].

Dado que é relevante o estudo do uso de sistemas de ASR para avaliação de fluência, este trabalho se difere dos demais por (1) apresentar uma abordagem aplicável a crianças sendo alfabetizadas em língua materna, (2) mostrar como o uso de alinhadores forçados trazem melhores resultados nesta tarefa e, (3) diferente de trabalhos anteriores, avaliar os resultados com uma quantidade expressiva de áudios avaliados manualmente.

3. Proposta desenvolvida

Um sistema de ASR clássico geralmente é utilizado para produzir a transcrição do áudio fornecido como entrada. No entanto, a utilização da transcrição para a avaliação da fluência pode mascarar a pronúncia incorreta de algumas palavras do texto de referência, bem como reconhecer palavras que não estão presentes no texto original, uma vez que o sistema busca sempre associar um conjunto de fonemas com a palavra mais provável. Embora esse modo de funcionamento seja adequado para muitas aplicações, a aproximação faz com que palavras pronunciadas incorretamente possam ser identificadas como corretas, encobrindo erros cometidos pelo falante. Trocar a sílaba tônica de uma determinada palavra, por exemplo, é um erro que deveria ser contabilizado. No entanto, como neste exemplo, a mudança na transcrição fonética da palavra se dá em apenas um fonema, o sistema de ASR tende a aproximar a pronúncia para a palavra correta, mascarando o erro de pronúncia. Omitir e adicionar plurais e conjugações de verbos são outros exemplos de erros que podem ser encobertos quando se leva em consideração apenas a transcrição.

Nesse contexto, este trabalho propõe o uso de algoritmos de alinhamento forçado em conjunto com o sistema de ASR. Os alinhadores forçados são capazes de alinhar temporalmente o que é pronunciado no áudio em relação a um texto de referência, de forma que apenas as palavras de interesse sejam reconhecidas e transcritas nos momentos es-

perados [Moreno et al. 1998, McAuliffe et al. 2017]. A partir da informação temporal, é possível determinar o intervalo de tempo entre início e fim da pronúncia de uma palavra, disponibilizando maiores informações relevantes para as métricas implementadas posteriormente. Além disso, a combinação da informação temporal com a transcrição bruta torna menos provável o encobrimento de erros de pronúncia mais simples e torna possível identificar erros como a omissão de frases ou erros de prosódia.

O processo de avaliação de fluência em leitura começa com a submissão dos arquivos de áudio e o texto referentes à leitura realizada pela criança à um *web service*. Nessa etapa, o *web service* faz as validações das entradas, gera um modelo léxico adaptado para ser capaz de converter os fonemas em apenas palavras presentes no texto de referência e as transporta para o alinhador. Na etapa do alinhamento, tem-se naturalmente como parte essencial da tarefa de alinhamento um sistema ASR. O modelo acústico não sofre nenhuma alteração quando um sistema de ASR é utilizado em conjunto com alinhadores, entretanto o modelo de linguagem é construído para que sejam reconhecidas apenas a sequência de palavras presente no texto de referência. Como resultado, fornece um conjunto de referências dos fonemas com trechos do texto de referência. Esse conjunto é posteriormente processado para gerar métricas quantitativas para avaliação de fluência (Seção 2.1). Uma visão geral do processo pode ser encontrada na Figura 1.

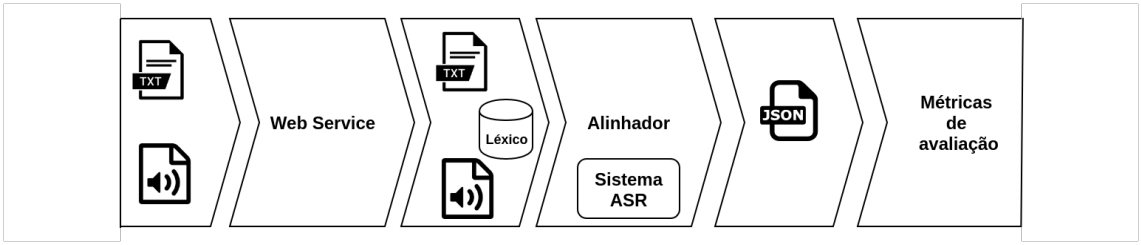

Figura 1. *Pipeline* para avaliação automática de fluência utilizando alinhador forçado

Um dos problemas oriundos da utilização de alinhadores forçados é determinar qual foi a última palavra lida do texto de referência. Mesmo que seja definido um tempo limite para uma gravação, não é possível garantir que a leitura gravada é correspondente a todo o texto podendo haver cortes na leitura durante a gravação do áudio. Isso faz com que as entradas não estejam sincronizadas, ou seja, o texto e o áudio não possuem as mesmas palavras que podem ser lidas fazendo com que o alinhamento de ambos possa não ser realizado corretamente. Assim, a detecção da última palavra pronunciada no áudio é de extrema importância, pois permite que o texto de referência seja cortado nessa palavra e que seja repassado para o alinhamento apenas a porcentagem do texto que foi de fato lida, aumentando a acurácia das métricas de fluência na etapa final do processamento.

Por conta desse problema, é acrescentado no *pipeline* de processamento um módulo auxiliar para identificação das palavras realmente pronunciadas e adaptação do texto (Figura 2). Para funcionamento desse módulo, é utilizado uma estrutura formal chamada autômato. Um autômato é constituído de um conjunto de possíveis estados, uma função de transição que mapeia as transições entre os estados e uma entrada de processamento. Nesse módulo, um autômato é construído com o texto de referência onde cada palavra do texto se torna uma transição entre estados, podendo a leitura acabar em qualquer parte do texto de referência. Assim, tendo como entrada de processamento o próprio

Figura 2. *Pipeline* para avaliação automática de fluência adicionando o módulo de verificação da última palavra pronunciada

áudio, com o auxílio do modelo acústico, o ASR percorre o áudio, realiza a identificação dos fonemas e verifica quais palavras são pronunciadas em sequência. Quando o áudio termina, a última palavra lida é determinada com base no último estado alcançado pelo autômato. Tendo a última palavra sido detectada, o *pipeline* de processamento segue para as etapas de Alinhamento e Métricas de Avaliação, porém agora com o texto contendo apenas as palavras que foram pronunciadas. O Algoritmo 1 apresenta o pseudocódigo do processo de construção do autômato do texto passado como referência (*word list*).

Algorithm 1: CONSTRUÇÃO DO AUTÔMATO **A** PARA CORTE DO TEXTO

Input : $word\_list$
Output: $\mathrm{A}(\Sigma, Q, \delta, e_0, F)$ para identificação da última palavra
begin
  $\Sigma \leftarrow \{word\_list[0], \cdots, word\_list[n-1]\}$, onde $n = len(word\_list)$
  $Q \leftarrow \{q_0, q_1, \cdots, q_n\}$
  $e_0 \leftarrow q_0$
  $F \leftarrow \{q_0, q_1, \cdots, q_n\}$
  for $i \leftarrow 0$ ***to*** $n-1$ **do**
    $\delta(q_i, word\_list[i]) \rightarrow q_{i+1}$
  end
end

4. Experimentos e Resultados

Os experimentos discutidos nesta seção consistem em uma análise comparativa entre as 3 abordagens para avaliação de fluência em leitura discutidos anteriormente: ASR, ALN e LSW. O **ASR** é o sistema de reconhecimento de fala padrão, o **ALN** é a utilização de alinhadores forçados em conjunto com sistemas ASR e o **LSW** é utilização do ALN em conjunto com o módulo de detecção de última palavra. Todas as abordagens utilizaram o mesmo modelo acústico. Além disso, as abordagens ASR e ALN utilizaram o mesmo modelo de linguagem construído com todas as palavras contidas no texto passado como referência, apenas a LSW teve o modelo de linguagem construído até a última palavra pronunciada.

Para a avaliação, foi utilizada uma base com aproximadamente 17.000 áudios com leituras realizadas por crianças gravadas dentro do ambiente escolar cedida pelo CAEd/UFJF (Centro de Políticas Públicas e Avaliação da Educação[1]) . Cada áudio possui

[1] `http://www.caed.ufjf.br/`

duração aproximada de 1 minuto de gravação contendo a leitura de um texto de referência pelo qual a leitura da criança seria avaliada. Além disso, cada um desses áudios foi avaliado manualmente por especialistas em avaliação de fluência contratados. A avaliação humana gerou um arquivo contendo a quantidade de palavras que não foram lidas ou foram lidas com erro pela criança, qual foi a última palavra lida e se o falante obedeceu as pausas de sentido. Esses dados são utilizados para avaliar o desempenho do classificador automático através da quantificação das métricas de QPL e QPC da avaliação manual em relação as encontradas automaticamente por cada abordagem. Com os valores de QPL e QPC, os leitores podem ser classificados em 2 grupos: fluente ou não-fluente. Os grupos foram definidos pelos especialistas da área no qual o grupo **fluente** consiste nos leitores que conseguem ler acima de 65 palavras corretas com 90% de precisão de leitura (QPC/QPL) e a classe **não-fluente** são todos os que não atingiram essa pontuação.

Foram utilizadas três métricas da literatura para avaliação: Precisão, Revocação e Acurácia. A **precisão** consiste na porcentagem de áudios classificados corretamente dentro de uma classe, a **revocação** consiste na porcentagem de áudios classificados corretamente entre todos que deveriam ser classificados naquela classe e a **acurácia** que mede a razão dos acertos do classificador, a qual consiste na quantidade de áudios classificados corretamente tanto como fluentes quanto não-fluente dentre todos os áudios. Ainda, foram medidos os tempos de processamentos para cada abordagem. Os experimentos foram realizados em uma máquina virtual *Intel(R) Xeon(R) CPU E5504 @ 2.00GHz 20GB*. A Tabela 1 apresenta os resultados obtidos. A primeira parte da tabela apresenta a comparação quantitativa entra cada abordagem e a avaliação manual, a segunda parte apresenta as métricas de avaliação e, por último, a terceira parte apresenta o tempo de processamento.

Tabela 1. Matriz de confusão com a classificação dos áudios. Leia: F como Fluente, NF como Não-Fluente , P como Precisão, R como Revocação, A como acurácia, T como tempo total em horas e M como Média em segundos.

		Manual		**Avaliação**			**Tempo**	
		F	NF	P(%)	R(%)	A(%)	T(h)	M(s)
ASR	F	473	2	99,58	6,52	60,21	489,95	$103,75 \pm 110,16$
	NF	7283	9798	59,09	99,98			
ALN	F	2749	52	98,14	37,89	73,27	334,36	$70,81 \pm 30,17$
	NF	4507	9798	68,38	99,47			
LSW	F	2792	52	98,17	38,48	**73,52**	**318,40**	$67,43 \pm 15,38$
	NF	4464	9746	68,59	99,47			
Total de áudios		**7256**	**9798**					

Analisando primeiramente a classe fluente, tanto a abordagem ALN quanto a LSW classificam aproximadamente 6 vezes mais áudios corretamente do que a ASR. Isso é refletido diretamente nas métricas de avaliação, visto que, embora o ASR avalie com uma precisão de 99,58% (aproximadamente 1,5pp maior que as demais), tem uma revocação muito menor que as outras abordagens. Nenhuma das abordagens consegue alcançar valores relativamente altos para revocação nessa classe, porém, vale ressaltar que todas elas possuem valores de precisão próximo de 100%. Isso se mostra relevante para avaliação da fluência, pois a alta presença de falsos positivos tem impactos muito mais negativos para

definições de intervenções ou políticas públicas para alfabetização do que a presença de falsos negativos. Após uma análise dos falsos positivos e negativos, verificou-se que esse grupo é formado principalmente por leituras perto do limitante de 65 palavras. O erro na classificação se dá por conta da qualidade de algumas gravações. Como o ambiente escolar geralmente não contém um espaço preparado acusticamente para gravações desse tipo, a presença de ruídos sonoros que distorcem as identificação dos fonemas pronunciados faz com que palavras corretas sejam classificadas como incorretas, diminuindo a revocação na classe de fluentes.

Para a classe não-fluente, o cenário é o contrário. O ASR classifica muitos áudios que eram pra ser fluentes como não-fluentes refletindo diretamente na precisão desta classe, a qual apresenta valores menores que as demais abordagens. Em termos da revocação da classe, todas as abordagens apresentam valores de revocação altos, acima de 99%. Entretanto, isso não se torna um dado tão relevante quanto a revocação na classe fluente, visto que a quantidade de áudios classificados como não-fluentes é maior e o classificador tende a jogar os áudios de classe fluente para a não-fluente.

Comparando a acurácia geral do classificador, a abordagem LSW apresenta a maior acurácia, com 73,52% seguido pela ALN, com 73,27% e, por último, o ASR com 60,21%. Com esses valores, mostra-se que o uso de alinhadores forçados realmente apresentam resultados satisfatórios para o cenário de avaliação de fluência. Entretanto, a diferença entre as abordagens ALN e LSW é muito baixa, ficando apenas 0,25pp de distância. Isso se dá porque a abordagem LSW consegue classificar mais áudios corretamente como fluentes, entretanto, para a base utilizada, essa quantidade foi muito pequena apresentando apenas 40 áudios a mais que a abordagem ALN. Por outro lado, para um processamento em larga escala, a abordagem LSW tende a apresentar resultados ainda mais satisfatórios, visto que classifica a classe fluente mais corretamente que a ALN, aumentando a revocação de áudios nessa classe.

Por último, analisa-se o tempo de processamento. Além do ASR apresentar resultados de acurácia pouco satisfatórios, apresenta o processamento mais lento levando em média 103s por áudio (quase 2x o tempo original do áudio). Dentre os áudios usados durante a avaliação, alguns possuem baixa qualidade acústica fazendo com que o ASR se perdesse durante o processo de transcrição. Para um segundo experimento retirando esse subconjunto, o tempo médio passou para 72s, ficando ainda assim maior que as demais abordagens. Já para a ALN e a LSW, que apresentaram valores de acurácia próximos, o tempo de processamento se distancia. Além da LSW apresentar uma média de processamento menor e um desvio padrão muito mais estável que a ALN, a abordagem gastou quase 17 horas a menos. Isso ocorre por conta do corte no texto realizado pelo módulo de detecção da última palavra, que encurta a quantidade de transições que o modelo de linguagem irá necessitar percorrer para avaliar um determinado áudio. Assim, pensando na avaliação em larga escala em todo território brasileiro, o tempo de processamento levado pela LSW será muito mais rápido e estável do a ALN.

5. Considerações finais e trabalhos futuros

O objetivo deste trabalho foi discutir como sistemas de reconhecimento de fala podem ser adaptados para avaliar a fluência em leitura de crianças. Nesse cenário, crianças ainda estão aprendendo a ler e, portanto, estão sujeitas a falhas frequentes de leitura, como sal-

tos de palavras ou erros de pronúncia. Sistemas ASR podem não ser eficazes se utilizados sozinhos, visto que tendem a encobrir esses erros e aproximar as pronúncias realizadas. Nessa visão, foi proposta a utilização de alinhadores temporais forçados para tentar contornar esse problema juntamente da utilização de um módulo para detecção da última palavra pronunciada.

Foi apresentada uma análise comparativa entre 3 abordagens diferentes: **ASR**, **ALN**, e **LSW** em uma base real com aproximadamente 17.000 áudios com gravações de avaliações de fluência de leitura aplicada em crianças. Os resultados mostraram que a abordagem LSW desenvolvida para esse projeto possui um tempo de processamento menor e mais estável, além de atingir uma melhor acurácia. Como contribuições desse trabalho, destaca-se a apresentação da abordagem para um cenário de aplicação ainda pouco explorado na literatura (avaliações de larga escala em oralidade) e a avaliação do trabalho com uma base real e com grande volume de áudios (~17000 áudios, ~285 horas).

Ainda é necessário explorar alguns refinamentos na abordagem para melhorar a revocação da classe fluente. Como o ambiente escolar não é preparado acusticamente para se realizar gravações com a finalidade de avaliação, existem áudios com ruídos sonoros como falas cruzadas e barulho ao fundo das gravações. Métodos computacionais para minimizar o impacto desses ruídos precisam ser investigados. Além disso, a fluência é apenas um dos aspectos para avaliação da alfabetização, sendo necessário, por exemplo, investigar métodos para auxiliar na avaliação da compreensão de textos de forma automática e aplicáveis em avaliações em larga escala.

Agradecimentos

Agradecemos ao CAEd/UFJF pelo financiamento do projeto e disponibilização da base de avaliação utilizada nos experimentos.

Referências

Campos, A. and Freitas, J. (2016). Reconhecimento automatico de fala (asr) e aquisição de segunda língua: Práticas de pronúncia do inglês no aplicativo móvel babbel. *Simpósio Internacional de Educação e Comunicação, Aracaju*.

Carchedi, L. C., Barrére, E., and Souza, J. (2018). 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.

Celestino, P. G. (2019). A oralidade infantil e desenvolvimento cognitivo à partir da prática docente. *Revista Internacional de apoyo a la inclusión, logopedia, sociedad y multiculturalidad*, 5(1).

Claus, F., Gamboa-Rosales, H., Petrick, R., Hain, H.-U., and Hoffmann, R. (2013). A survey about asr for children.

Demenko, G., Wagner, A., and Cylwik, N. (2010). The use of speech technology in foreign language pronunciation training. *Archives of Acoustics*, 35(3):309–329.

Eskenazi, M. (1996). Detection of foreign speakers' pronunciation errors for second language training-preliminary results. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1465–1468. IEEE.

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.

Fuchs, L. S., Fuchs, D., Hosp, M. K., and Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific studies of reading*, 5(3):239–256.

Hoy, M. B. (2018). Alexa, siri, cortana, and more: An introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88.

Hudson, R. F., Lane, H. B., and Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher*, 58(8):702–714.

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N., Jiang, Q.-M., Sainath, T. N., Senior, A., Beaufays, F., and Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Litman, D., Strik, H., and Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 15(3):294–309.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502.

MEC (2018). Base nacional comum curricular. `http://basenacionalcomum.mec.gov.br/wp-content/uploads/2018/02/bncc-20dez-site.pdf`. Acessado: 30-01-2019.

Moreno, P. J., Joerg, C., Thong, J.-M. V., and Glickman, O. (1998). A recursive algorithm for the forced alignment of very long audio segments. In *Fifth International Conference on Spoken Language Processing*.

Neri, A., Cucchiarini, C., and Strik, H. (2006). Asr corrective feedback on pronunciation: Does it really work?

Neri, A., Cucchiarini, C., and Strik, W. (2003). Automatic speech recognition for second language learning: how and why it actually works. In *Proc. ICPhS*, pages 1157–1160.

Soares, E., Carchedi, L. C., Gomes Jr, J., Barrére, E., and Souza, J. (2018). 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.

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *Calico Journal*, 28(3):744.

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

Yu, D. and Deng, L. (2016). *AUTOMATIC SPEECH RECOGNITION*. Springer.