Avaliação de modelos para reconhecimento automático de fala aplicados para identificação da qualidade de leituras em voz alta de narrativas breves

André Luiz Vasconcelos Ferreira[1], **Cristiano Nascimento Silva**[1], **Elias Cyrino de Assis**[1], **Jairo Francisco de Souza**[1,2]

[1]LApIC Research Group – UFJF – Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF) – Juiz de Fora, MG – Brasil

{cristiano.nascimento,andre.vasconcelos,elias.cyrino,jairo.souza}@ice.ufjf.br

Abstract. *Advances in the area of automatic speech recognition (ASR) have allowed the emergence of innovative solutions in the area of Informatics in Education, especially in the domain of literacy assessment. Its use for child speech recognition, however, still brings challenges, and studies that analyze new technologies in this application domain are lacking. This article presents a comparison between two ASR technologies in the context of children's speech for the automatic assessment of reading fluency: a supervised and a self-supervised approach. 59 audios of children's readings aloud were used. Wav2Vec2 together with a language model showed substantially better results than the other models in relation to the word error rate.*

Resumo. *Os avanços na área de reconhecimento automático de fala (ASR) tem permitido o surgimento de soluções inovadoras na área de Informática na Educação, especialmente no domínio de avaliação da alfabetização. O seu uso para reconhecimento de fala infantil, contudo, ainda traz desafios e faltam trabalhos que analisem novas tecnologias neste domínio de aplicação. Este trabalho apresenta uma comparação entre duas tecnologias de ASR no contexto de fala de crianças para a avaliação automática de fluência de leitura: uma abordagem supervisionada e uma abordagem auto-supervisionada. Foram utilizados 59 áudios de leituras de crianças. O Wav2Vec2 em conjunto com um modelo de língua apresentou resultados substancialmente melhores que os demais modelos em relação à taxa de erro de palavras.*

1. Introdução

As avaliações em larga escala têm sido utilizadas nas últimas décadas como instrumentos direcionadores de políticas públicas e indicadores de qualidade de ensino [Sorgatto et al. 2021]. Como explicado em [Bauer et al. 2015], entre as vantagens deste tipo de avaliação na área de educação, pode-se citar o fato de que ela responsabiliza instituições e profissionais de ensino pelos resultados alcançados, o que motiva-os na melhoria da execução de seus trabalhos, estabelece uma cultura de transparência dos serviços públicos que as aplicam, auxiliam pais na escolha de quais estabelecimentos merecem confiança para o aprendizado de seus filhos, desafiam estudantes na

DOI: 10.5753/sbie.2022.224744

busca por melhores resultados e auxiliam na correção de currículos escolares inadequados. Aliado a isso, a automatização dessas avaliações se mostra relevante uma vez que diminui custos, o tempo de execução do processo de avaliação [Carchedi et al. 2018, Carchedi et al. 2021] e o impacto da subjetividade do avaliador humano na correção das atividades [Bauer et al. 2015].

A avaliação automática de itens para identificação da qualidade de leitura oral de crianças é desafiadora do ponto de vista computacional, porém de interesse para viabilizar avaliações em larga escala sobre a alfabetização infantil. Dentre os desafios, está a dificuldade da máquina identificar erros e padrões de leitura que um avaliador humano conseguiria [Beck et al. 2004]. Outro fator que dificulta a automatização desse tipo de avaliação é o fato desta ter como objetivo a leitura de crianças, pois elas são inconsistentes na articulação de palavras e, com a habilidade de leitura ainda em desenvolvimento, modelar sistemas computacionais que entendam-nas se torna ainda mais difícil [Sabu and Rao 2018].

Entre as tecnologias de reconhecimento de fala, destaca-se para este trabalho o uso das *Time delay neural network* (TDNN) que podem ser encontradas em *toolkits* como o Kaldi [Povey et al. 2011]. Neste tipo de abordagem, bases de dados com áudios e transcrições adequadamente preparadas para treinamento de sistemas de reconhecimento de fala são utilizadas. Contudo, é custosa a criação destas bases. Recentemente, o uso de abordagens de aprendizado auto-supervisionado para aprender representações da fala em áudios não rotulados tem diminuído o custo de preparação de bases de dados de treinamento e apresentado resultados promissores em reconhecimento de fala. O uso de ambas as tecnologias no reconhecimento de fala espontânea de crianças tem sido alvo de pesquisas recentes [Jain et al. 2022, Bhardwaj et al. 2020]. O presente trabalho, por sua vez, visa comparar esses dois tipos de abordagens para a tarefa de avaliação automática de fluência em leitura de crianças em fase de alfabetização de língua portuguesa. Para isso, foram treinados dois modelos (um modelo TDNN e um modelo auto-supervisionado chamado Wav2Vec2) e avaliados com 59 áudios de leituras de itens narrativos breves utilizados em avaliações em larga escala aplicados em escolas públicas brasileiras.

Este artigo está organizado como se segue. Na seção 2 são apresentados os trabalhos relacionados e os conceitos básicos em relação às redes TDNN e o *framework* Wav2Vec2. Em seguida, a Seção 3 contém a descrição da base de dados utilizada neste estudos e os procedimentos de treinamento dos modelos. A descrição dos experimentos e análise dos resultados são apresentadas na Seção 4. Por fim, a Seção 5 apresenta as considerações finais e trabalhos futuros.

2. Fundamentação teórica

O reconhecimento de fala na área de Educação tem sido utilizado com resultados importantes há muitos anos. Esta seção apresenta os trabalhos relacionados, destacando o uso da tecnologia para avaliação de leitura, e termina detalhando as características das duas tecnologias de reconhecimento de fala que são alvos do presente estudo.

2.1. Trabalhos Relacionados

Um dos primeiros trabalhos a utilizar tecnologias de reconhecimento de fala no contexto educacional é o de [Bernstein et al. 1990], no qual estudantes japoneses tinham suas habilidades em língua inglesa avaliadas por um sistema que processava áudios de leitura.

Posteriormente, pesquisas demonstraram a efetividade do uso de reconhecimento de fala no processo de melhoria da fluência de leitura [Mostow et al. 2003, Beck et al. 2004, Poulsen et al. 2007, Reeder et al. 2007]. Entre as primeiras formas de se realizar esse tipo de avaliação, pode-se citar o cálculo do tempo entre a leitura correta de duas palavras da avaliação [Mostow et al. 2003]. Esse tipo de teste indica o esforço do leitor em decodificar as palavras com as quais se depara em uma frase [Bolanos et al. 2013].

Outra abordagem para uso de reconhecimento de fala na avaliação de fluência de leitura é a que consiste na avaliação da alfabetização por meio da leitura de palavras isoladas [Black et al. 2007, Black et al. 2008, Black et al. 2010]. Com isso, é possível a detecção de características da leitura que definem o nível de fluência do leitor, como a presença de sussurros, hesitações ou palavras lidas com a entonação de perguntas [Bolaños et al. 2011]. Posteriormente, o método de avaliação foi aprimorado com a implementação de recursos que permitiam falsos começos, repetições e deleções de palavras durante as leituras [Duchateau et al. 2007]. Com o tempo, surgiram trabalhos que, além de realizarem a avaliação de fluência de leitura com listas de palavras, também o fizeram com passagens de texto [Zechner et al. 2009, Yilmaz et al. 2014].

Ainda foram desenvolvidas pesquisas cuja contribuição era a modelagem de sistemas de reconhecimento de fala tendo como unidade de reconhecimento subpalavras, como sílabas, ao invés de palavras [Hagen et al. 2004, Hagen and Pellom 2005, Hagen et al. 2007]. A motivação para esse tipo de abordagem vem da noção de que muitas dificuldades de leitura estão relacionadas à pronúncia de partes da palavra ao invés dela como um todo [Bolaños et al. 2011]. Em seguida houve o uso de Máquinas de Vetor Suporte na reclassificação de unidades de subpalavras [Bolanos 2008], o que melhorou os resultados das pesquisas anteriores uma vez que houveram menores taxas de erro de palavras para as mesmas bases de dados.

Segundo Sabu *et al* [Sabu et al. 2017], tradicionalmente os sistemas de reconhecimento de fala utilizados para avaliação automática de leitura são baseados em Modelos Ocultos de Markov (HMM) [Duchateau et al. 2007] com Modelos de Mistura Gaussiana (GMM), mas com o tempo modelos baseados em Redes Neurais Profundas se mostraram mais eficientes [Metallinou and Cheng 2014, Cheng et al. 2015, Tao et al. 2016]. Contudo, o problema de criação de modelos adequados para reconhecimento de fala de criança persiste, o que influencia o avanço de pesquisas de aplicação de IA na área de alfabetização. Dentre os problemas existentes está a necessidade de amplo volume de dados rotulados para treinamento adequado de modelos que sejam capazes de identificar variações de fala e regionalismos. Recentemente, abordagens com treinamento auto-supervisionado têm permitido o surgimento de modelos pré-treinados que podem ser especializados para determinadas tarefas utilizando conjuntos mais reduzidos de áudios, como é o caso do Wav2Vec2 [Yu et al. 2021, Fan et al. 2021, Jain et al. 2022]. Contudo, essa abordagem ainda não foi explorada no contexto de avaliação de fluência. A contribuição desse trabalho, portanto, está na comparação entre o uso de uma abordagem tradicional (usando uma rede TDNN) e uma abordagem auto-supervisionada (treinando uma rede com Wav2Vec2) para avaliação de fluência de crianças em fase de alfabetização.

2.2. Arquiteturas para reconhecimento de fala

Existem diversas arquiteturas de redes neurais criadas para reconhecimento de fala (ASR), as quais duas são abordadas nesta seção: a TDNN e o Wav2Vec2. Redes usadas para

ASR geralmente têm como entrada um áudio segmentado em *frames*, o que permite a identificação de fones correspondentes a estes, resultando então na transcrição da mídia acústica. Os fones são unidades acústicas que representam sons específicos. A identificação dos fones, na verdade, é dada por meio de uma distribuição de probabilidades entre eles. Aquele cujo resultado possui maior probabilidade segundo os cálculos da rede neural é tido como o fone correspondente ao *frame*.

A arquitetura *Time Delay Neural Network* (TDNN) [Tchistiakova 2019] é uma arquitetura multicamadas que pode ser usada em diversos tipos de problemas, como classificação de imagens, mas também muito usada para processar fala, sendo encontrada em *toolkits* como o Kaldi [Povey et al. 2011]. Neste contexto, esta rede é usada para classificar fones com *shift-invariance* (isto é, a rede evita ou não necessita que tempos de início e fim de cada fone sejam determinados nos áudios antes da classificação) e modelar contextos (isto é, cada neurônio em cada camada recebe entrada não apenas das ativações do nível anterior, mas também de um padrão de saída e seu contexto).

Os áudios de entrada da TDNN são representados por uma série de *features* acústicas guardadas em vetores $x \in \mathbb{R}^m$, em que m é o número de *features*. Cada vetor x representa um *frame* do áudio de entrada. Dessa forma, a entrada da TDNN é um conjunto de vetores dado por $X \in \mathbb{R}^{m \times t}$, em que m é o número de *features* do áudio e t é o número de *frames* em que o áudio foi segmentado. Além disso, outro elemento importante de uma TDNN é a matriz treinável de pesos, dada por $W \in \mathbb{R}^{m \times l}$, em que $l < t$. A matriz passa pelos *frames* do áudio e, por meio de operações convolucionais, gera as saídas da rede neural.

Após gerar uma saída, a matriz treinável se locomove s vezes sobre o conjunto de *frames* para criar a próxima saída. O número de saídas o, então, é dado pela fórmula $o = \left\lfloor \frac{t-l+2b}{s} + 1 \right\rfloor$. As saídas da TDNN são calculadas por meio de operações convolucionais, ou seja, operações cujo resultado é a soma das multiplicações dos pesos da matriz treinável com os valores das *features* do áudio de entrada da rede neural. O valor obtido dessas operações é então somado a um viés b e o resultado disso é submetido a uma função não linear de ativação ϕ. A função ϕ varia de acordo com a implementação da TDNN. Assim, considerando um conjunto Z de saídas da rede neural, o valor de uma saída $z_q \in Z$, com $q \leq o$ e $q \in \mathbb{N}$, é dado por $z_q = \phi(\sum_{i=0}^{m} \sum_{j=0}^{l} w_{i,j} x_{i,j} + b)$, em que $w_{i,j}$ é o valor na matriz treinável de pesos na coordenada (i, j) e $x_{i,j}$ é o valor da *feature* no vetor que representa o *frame* visitado pela matriz treinável de pesos. O processo descrito de entrada e saída da TDNN pode ocorrer em diversas camadas, ou seja, a saída de uma camada é utilizada como entrada de uma outra camada do processo e assim por diante.

Por sua vez, o Wav2Vec2 se trata de um modelo pré-treinado auto-supervisionado [Vaessen and van Leeuwen 2021]. Para entender essa abordagem, é importante compreender modelos de aprendizagem supervisionada e não supervisionadas. O primeiro tipo consiste em treinar o modelo por meio de *labels*. Por exemplo, se é desejada a criação de um sistema que reconheça um gato em uma imagem, diversas figuras com e sem o animal, marcadas respectivamente com os *labels* "gato" e "não há gato", serão apresentadas ao modelo de aprendizagem de máquina de forma que o mesmo aprenda a fazer o reconhecimento. O segundo tipo consiste em treinar o modelo sem os *labels*, fazendo-o identificar padrões a partir apenas dos dados de treinamento. No exemplo do sistema que reconhece gatos em imagens, o modelo seria treinado apenas com figuras do animal até

que padrões nas imagens fossem encontrados de forma a resolver o objetivo. O modelo de auto-aprendizado consiste em sistemas que aprendem a identificar partes de um dado a partir de outras partes do mesmo dado. Isso é feito a partir da geração de *labels* advindos de padrões identificados pelo sistema nos dados que recebe, caracterizando assim um modelo não supervisionado que se transforma em um modelo supervisionado.

Para efetuar o reconhecimento de fala, o Wav2vec2 é usado com duas etapas. A primeira é o pré-treino com um grande número de áudios não rotulados. A segunda consiste em um processo de treinamento não supervisionado seguido de uma especialização do modelo (*fine tuning*) [Jain et al. 2022]. O Wav2vec2 recebe uma entrada de áudio χ. Essa entrada então é processada com um codificador $f : \chi \mapsto \zeta$ que divide χ em saídas $z_0, ..., z_\tau$ que representam *frames* do áudio, em que τ é o número de espaços de tempo em que o áudio é dividido e $z_i \in \zeta$ com $0 \leq i \leq \tau$ e $i \in \mathbb{Z}$. Dessas saídas, são produzidos através de um transformador $g : \zeta \mapsto \varsigma$ um conjunto de valores ς dados por $c_0, ..., c_\tau$, bem como, através de um quantizador $\zeta \mapsto \varrho$, um conjunto de valores dados por q_t [Baevski et al. 2020]. Os valores dos conjuntos ς e ϱ são então utilizados no aprendizado do reconhecimento de fala durante o pré-treino. Os valores de ς são representações contextualizadas dos valores de ζ, ou seja, valores que o sistema deve aprender a reconhecer para um conjunto de entradas de treinamento. Quando o sistema estiver treinado o suficiente para reconhecer os valores de ς, ele estará pronto para ser usado fora do contexto do treinamento. Por sua vez, ϱ se trata de um conjunto de representações de ζ que, dependendo da forma como são combinadas, permitem que o sistema reconheça os valores do conjunto ς. O aprendizado do reconhecimento de fala do Wav2vec2, portanto, se dá quando o sistema utiliza corretamente os valores de ϱ na obtenção dos valores de ς. Para que esse aprendizado ocorra, são feitos cálculos de probabilidade que comparam a similaridade de cada valor c de ς com todos os valores q de ϱ. Desses cálculos, as maiores probabilidades resultantes indicam quais valores q se relacionam ao valor c que o sistema está aprendendo a identificar [Baevski et al. 2020, Jain et al. 2022].

Sistemas que usam Wav2vec2 para ASR podem ser considerados sistemas ponta-a-ponta (*end-to-end*), uma vez que a rede é capaz de aprender os processos de segmentação de fones e classificação, enquanto redes TDNN são mais especializadas e, por isso, exigem etapas adicionais para o reconhecimento de fala. Embora redes TDNN e similares possuem bons resultados para reconhecimento de fala, elas são dependentes de bases de treinamento bem estruturadas e de razoável volume[1] para que consigam gerar resultados satisfatórios. A criação dessas bases, contudo, é limitante em muitos projetos que necessitam de modelos específicos para ASR. O Wav2vec2, por sua vez, permite o uso de bases de dados menores para o processo de *fine-tuning*, o que pode trazer resultados satisfatórios para algumas tarefas de ASR, mas com um menor esforço.

3. Materiais e método

Os experimentos deste trabalho consistem em comparar as transcrições geradas a partir de um modelo supervisionado TDNN e um modelo de auto-supervisionamento (Wav2Vec2). Foram utilizados 59 áudios de leituras de dois textos narrativos com 250 palavras e diferente nível de complexidade (32 leituras de um texto e 27 de outro). As leituras foram feitas por crianças do 2º ano de colégios públicos do Maranhão. Estes 59 áudios foram

[1] Alguns especialistas recomendam o uso de pelo menos 100 horas de fala.

selecionados por possuírem uma qualidade de leitura mediana ou boa, ou seja, seus leitores conseguiram falar com clareza e identificar a maioria das palavras dos textos, além de boa qualidade de gravação, isto é, presença de apenas um falante ao longo do áudio e sem ruídos. Dessa forma, os modelos de reconhecimento de fala possuem maior chance de retornar seus melhores resultados. O texto retornado por cada modelo é comparado com o texto de referência daquele áudio e, assim, pode-se verificar a qualidade da transcrição em relação à referência. Os passos dos experimentos são discutidos nas subseções seguintes.

3.1. Predição com TDNN

Para treinar a rede TDNN foi utilizado o *toolkit* Kaldi, atualmente um dos *toolkits* gratuitos mais populares para reconhecimento de fala. O modelo foi treinado com 80 horas de áudios de leituras infantis que não fazem parte do conjunto de 59 áudios citados anteriormente. Dessas 80 horas, foram incluídas também 15 minutos de treinamento com áudios de leitura ruins para atingir uma melhor generalização do resultado do treinamento. Dos áudios utilizados, 80% foram para treinamento e 20% para validação com o intuito de treinar um modelo HMM-GMM (*Hidden Markov Models-Gaussian Mixture models*) de trifones com SAT (*speaker adaptive training)*. Este, por sua vez, foi usado no treinamento da TDNN, a qual contém 7 camadas com 384 neurônios, uma camada de saída com 512 neurônios, e uma função objetivo LF-MMI (*Lattice-free Maximum Mutual Information)*. A rede foi treinada por 25 épocas. Para atingir melhores resultados no reconhecimento de cada texto, foi criado um léxico e um modelo de língua para cada texto. Assim, o reconhecedor faz uso do modelo acústico treinado mas tende a escolher palavras presentes no texto alvo, o que permite resultados mais acurados.

3.2. Predição com Wa2Vec2

Por sua vez, o modelo reconhecdor que usa a rede Wav2Vec2 foi gerado através de um *fine-tuning* do modelo Wav2Vec2 pré-treinado para 50 idiomas. O *fine-tuning* visa treinar o modelo para a tarefa de reconhecimento de fala em língua portuguesa. Por limitações de *hardware*, optou-se por utilizar um modelo disponibilizado pela comunidade científica [Junior et al. 2021], o qual foi treinado com aproximadamente 290 horas de áudios na língua portuguesa acompanhados de suas transcrições. O treinamento durou 20 épocas.

Embora este modelo tenha sido treinado com um conjunto maior de áudios que o modelo TDNN, ressalta-se que foram usados apenas áudios de falantes adultos, o que poderia ser um dificultador para o modelo ao ser utilizado em áudios de crianças. Neste sentido, a avaliação desse modelo se torna relevante para entender sua capacidade de generalização para uso em avaliação de áudios infantis, uma vez que ainda há uma carência de bases gratuitas de áudios em português de leituras de crianças para treinamento de modelos de reconhecimento de fala.

Esse tipo de modelo é capaz de reconhecer as palavras sem o uso de um léxico, como o TDNN. Contudo, modelos de língua podem ser utilizados para auxiliar a predição. Assim, para aumentar a qualidade da transcrição do modelo, foi produzido um modelo de língua para cada texto e este foi opcionalmente utilizado após o reconhecimento dos fones.

3.3. Modelo de língua

O modelo de língua representa uma distribuição de probabilidades de ordens de palavras, chamadas de gramas. Por exemplo, se o modelo de língua é baseado em bigramas, este possui a probabilidade de ocorrência das ordens de duas palavras possíveis de um texto.

A Tabela 1 apresenta um trecho de um dos modelos de língua baseado em bigramas utilizado neste estudo. No exemplo do bigrama, é possível observar sequências de duas palavras possíveis de serem pronunciadas em áudios de leitura do texto T1. Junto a essas sequências de palavras, é definida uma pontuação que representa a probabilidade de ocorrência da sequência. Os valores são armazenamentos em logaritmo, então quanto mais próxima de zero, maior a probabilidade de ocorrência.

Tabela 1. Trecho do modelo de língua do texto T1 (formato ARPA)

2-grams:	
-1.1 a <unk>	-1.1 acha </s>
-1.1 agora <unk>	-1.1 agora </s>
-1.1 amigo <unk>	-1.1 amigo </s>
...	...
-1.1 <unk> a	-1.1 acha bonito
-1.1 <unk> amigo	-1.1 agora foi
-1.1 <unk> agora	-1.1 amigo é
...	...

A presença do <unk> prevê as aglutinações de sons desconhecidos após ou antes da identificação de alguma das palavras. Já com o bigrama "*palavra* </s>" é permitido que qualquer palavra presente no modelo possa ser marcada como fim de frase, permitindo que uma palavra em qualquer posição do texto possa ser reconhecida como palavra final, já que nem sempre ocorre a leitura do texto completo. Os altos pesos (valores -1.1) foram adotados para influenciar o reconhecedor de fala a encontrar palavras presentes no modelo de língua. Vale notar que o modelo de língua é usado pelo reconhecedor de fala após a identificação dos fones de cada *frame* do áudio e é usado para desambiguar um conjunto de palavras candidatas para uma sequência de *frames* do áudio. Assim, o uso do modelo de língua não força o reconhecedor a encontrar apenas palavras presentes nele, mas influencia na predição ao adicionar um peso nestas palavras.

4. Experimentos e resultados

Foram realizados experimentos com três abordagens: TDNN, Wav2Vec2 (sem uso do modelo de língua) e Wav2Vec2-lm (com modelo de língua). O modelo de língua usado no Wav2Vec2 é o mesmo usado no TDNN. O texto da transcrição de cada modelo foi alinhado com o texto de referência utilizando para produzir as medidas WER (*Word Error Rate*) e PO-WER (*Phonetically-Oriented WER*). Essas medidas são geradas à partir de um alinhamento entre o texto (chamado de hipótese ou HYP) e uma referência (REF). Para o WER são alinhadas cada palavra do texto, enquanto o PO-WER realiza um alinhamento de fonemas após um alinhamento das palavras, reduzindo a contagem de erros gerados por homofonia [Ruiz and Federico 2015]. A Tabela 2 mostra a diferença entre o alinhamento gerado pelo WER e pelo PO-WER, onde as inserções, substituições, remoções e palavras corretas são denotadas pelas letras I, S, R e C, respectivamente.

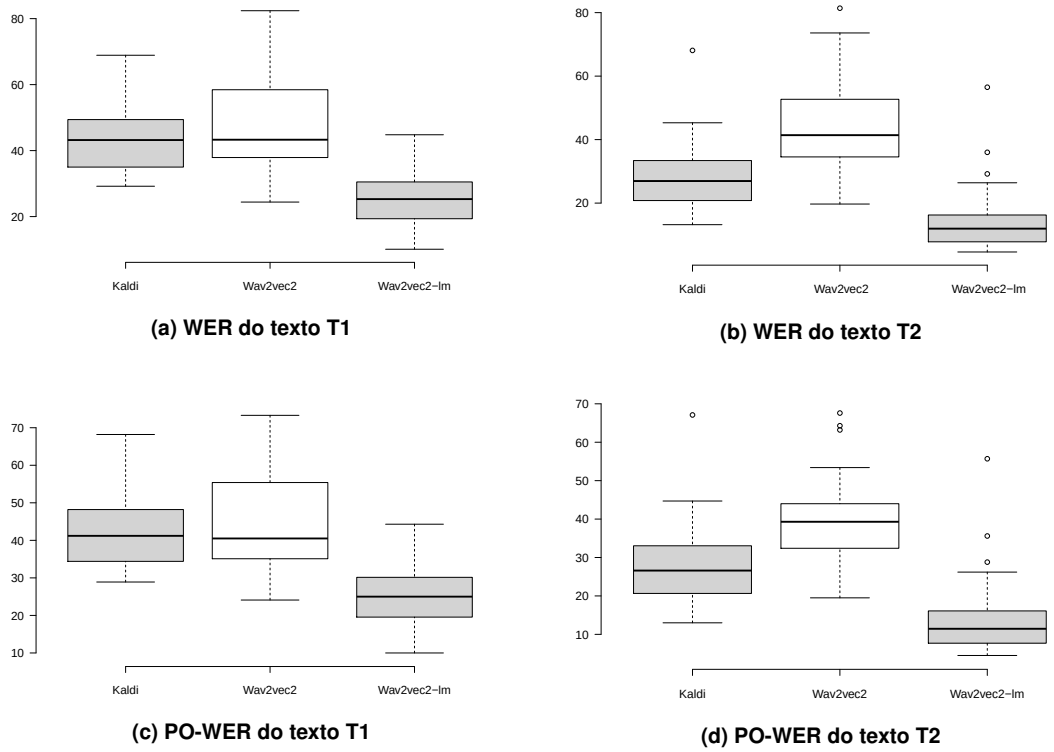
As métricas WER e PO-WER foram produzidas, respectivamente, à partir do alinhamento do Sclite[2] e do código dos autores do PO-WER[3]. Os dois textos utilizados estão

[2]https://github.com/usnistgov/SCTK
[3]https://github.com/NickRuiz/power-asr

Tabela 2. Exemplo de alinhamento gerado pelo WER e PO-WER

Resultado do WER: 4 erros								
REF:		carolina		teste	do	copo		demais
HYP:	glo	rinha		teste	do	copo	de	mais
Eval:	I	S		C	C	C	I	S
Resultado do PO-WER: 2 erros								
REF:	carolina		teste	do	copo	demais		
HYP:	glo	rinha	teste	do	copo	de	mais	
Eval:	S		C	C	C	S		

identificados neste artigo como T1 e T2. O resultado das métricas para os três modelos está apresentado na Figura 1.



(a) WER do texto T1

(b) WER do texto T2

(c) PO-WER do texto T1

(d) PO-WER do texto T2

Figura 1. Comparação da distribuição das métricas de erro entre os três modelos

Os resultados para o texto T1 são levemente piores do que para T2. Ressalta-se que os textos possuem diferente nível de dificuldade, o que gera também maior hesitação da criança ao longo da fala, gerando confusão durante a predição dos modelos. Contudo, as três estratégias de reconhecimento de fala tiveram comportamento parecido nos dois textos, onde o Wav2Vec2 sem modelo de língua gerou resultados piores e o Wav2Vec2 com modelo de língua conseguiu superar o TDNN. Reforça-se que o modelo TDNN foi treinado com áudios de crianças, enquanto o modelo Wav2Vec2 foi treinado com áudios de adultos. Assim, o Wav2Vec2, quando acompanhado de um modelo de língua especializado para o texto de referência, consegue produzir resultados melhores do que a aborda-

gem supervisionada usando TDNN. O uso de modelo de língua se mostrou fundamental para esses resultados. Neste sentido, é importante reforçar que, no contexto desse trabalho, o aluno é instruído a ler um texto conhecido. Dessa forma, é possível criar um modelo de língua que representa a gramática daquele texto, induzindo o reconhecedor de fala a encontrar palavras na sequência descrita na gramática. Para outras tarefas de reconhecimento, contudo, não é possível usar um modelo de língua com a distribuição de probabilidades adotada neste trabalho, como, por exemplo, na tarefa de reconhecimento de fala espontânea (ou seja, onde não há conhecimento prévio do que será falado).

O valor médio das métricas para cada texto está apresentada na Tabela 3, em que percebe-se que o Wav2Vec2-lm alcançou um WER duas vezes melhor que o TDNN (isto é, metade do valor).

Tabela 3. WER médio das transcrições de cada abordagem

	TDNN		**Wav2Vec2**		**Wav2Vec2-lm**	
	WER	PO-WER	WER	PO-WER	WER	PO-WER
T1	44,0%	42,7%	47,6%	43,6%	24,7%	25,1%
T2	28,3%	28,0%	45,1%	40,8%	14,8%	14,6%

Entre os motivos para o Wav2vec2 sem modelo de língua possuir um maior WER está o fato de muitos fones serem identificados, mas o modelo apresentar uma dificuldade na formação de palavras por conta da dificuldade do falante. Como são crianças em processo de alfabetização, a hesitação na fala, as pausas ao longo da pronúncia de palavras e os falsos começos na leitura de cada palavra confundem o modelo. Assim, ao formar palavras, é comum a produção de transcritos com palavras erradas mas próximas às esperadas. Por exemplo, a leitura de "a morada é perto do lago", por conta de leituras sem a devida fluência, com muitas pausas entre sílabas e com algumas tentativas de leitura (repetições de sílabas) podem gerar trascritos como "amor nada amornada esperto dodô lado". O efeito disso no cálculo do WER é a presença de muitas substituições, remoções e inclusões. Com o uso do modelo de língua no Wav2Vec2-lm e do TDNN, é aumentada a probabilidade de escolha de palavras presentes no modelo, influenciando na escolha de "do do" ao invés de "dodô", "a morada" ao invés de "amor nada", por exemplo. Por esse motivo, o PO-WER teve valores melhores que o WER na abordagem com o Wav2Vec2 mas obteve valores ligeiramente parecidos com o WER no Wav2Vec2-lm.

Verificou-se que o tamanho dos textos não está diretamente relacionado ao WER de cada áudio. A Figura 2 representa a relação do número de palavras lidas em áudios dos textos T1 e T2 (eixo X) e o WER do áudio (eixo Y) para o modelo de melhor resultado (Wav2Vec-lm). Como pode ser visto, não há um padrão que relaciona o número de palavras lidas com o WER da transcrição. Isso denotada que o padrão de WER entre 10% e 30% ocorre independente do quanto a criança conseguir ler e esse valor deve ser considerado para aplicações na área de Educação que vão fazer uso desse tipo de transcrição.

5. Considerações finais

A popularização de modelos e sistemas para reconhecimento de fala tem permitido o surgimento de muitas aplicações em Informática na Educação. Muitas das tecnologias

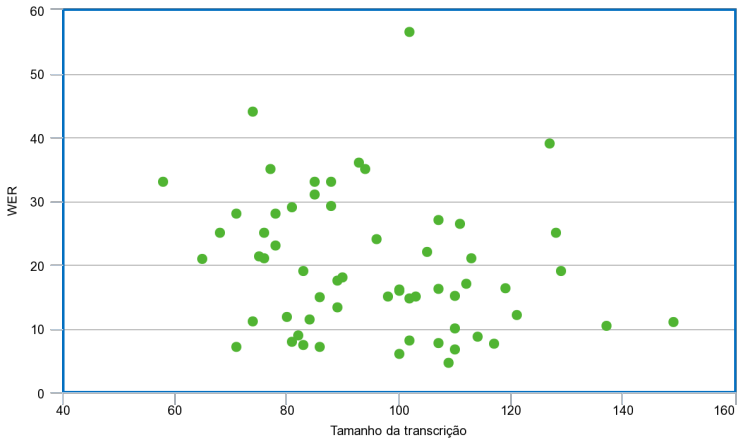


Figura 2. Diagrama que relaciona o tamanho da transcrição com o *WER*

criadas, contudo, são avaliadas com *benchmarks* conhecidos na área de Processamento de Linguagem Natural, os quais são comumente formados por áudios jornalísticos ou programas televisivos. Neste trabalho, avaliamos duas soluções para reconhecimento de fala com o objetivo de verificar o seu comportamento em áudios de crianças, o qual é público de interesse em muitas propostas de soluções inteligentes em Informática na Educação.

Foram realizados três experimentos com áudios de crianças com leituras em voz alta de duas histórias curtas. A taxa de erro de palavras ao comparar o texto gerado pelos modelos com o texto de referência demonstrou que o modelo Wav2Vec2-lm, mesmo que treinado com áudios com falas de adultos, conseguiu superar o modelo TDNN treinado com falas de crianças lendo outras histórias curtas. Os resultados apresentados neste estudo auxiliam pesquisadores e desenvolvedores na escolha de tecnologias para processar áudios de crianças. Uma vez que apresenta-se um modelo com menor WER, tem-se um reconhecimento mais próximo da fala e, com isso, a matriz de emissão é mais confiável para aplicações de avaliação de fluência, como verificação de prosódia, erros de leitura, entre outras características da fala que sujeitas à avaliação educacional.

Não existe pesquisa definitiva e os resultados desse estudo não podem ser considerados conclusivos. Neste trabalho os modelos foram avaliados com um conjunto pequeno de áudios, com um vocabulário controlado e leitura conhecida, além de pouca variação de leituras (dois textos). Os resultados demonstram uma tendência de escolha do uso do Wav2Vec2-lm, mas são necessários novos estudos com um volume maior de áudios para treinamento e avaliação, além de um conjunto mais diverso de textos e de falantes. Embora demande considerável esforço e custo, a criação de bases públicas de áudios com leituras de crianças em fase de alfabetização em língua portuguesa podem impulsionar pesquisas nesta área.

Os resultados encontrados, embora avaliados para a tarefa de *speech-to-text*, indicam que o modelo Wav2Vec2-lm possui maior acerto na predição de fones para cada *frame* do áudio, uma vez que sua comparação com o modelo TDNN apresentou melhor resultado. Contudo, o uso desses modelos pode apresentar resultados distintos em outras tarefas de reconhecimento de fala para áudios de crianças e necessitam de nova investigação.

Referências

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Bauer, A., Alavarse, O. M., and Oliveira, R. P. d. (2015). Avaliações em larga escala: uma sistematização do debate. *Educação e Pesquisa*, 41:1367–1384.

Beck, J. E., Jia, P., and Mostow, J. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. *Technology Instruction Cognition and Learning*, 2:61–82.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., and Weintraub, M. (1990). Automatic evaluation and training in english pronunciation. In *First International Conference on Spoken Language Processing*.

Bhardwaj, V., Kadyan, V., et al. (2020). Deep neural network trained punjabi children speech recognition system using kaldi toolkit. In *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, pages 374–378. IEEE.

Black, M., Tepperman, J., Lee, S., and Narayanan, S. S. (2008). Estimation of children's reading ability by fusion of automatic pronunciation verification and fluency detection. In *Ninth Annual Conference of the International Speech Communication Association*.

Black, M., Tepperman, J., Lee, S., Price, P., and Narayanan, S. S. (2007). 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*.

Black, M. P., Tepperman, J., and Narayanan, S. S. (2010). Automatic prediction of children's reading ability for high-level literacy assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):1015–1028.

Bolanos, D. (2008). *Advances in the application of support vector machines for continuous automatic speech recognition*. PhD thesis, Ph. D. thesis, Computer Science Department, Universidad Autonoma de Madrid.

Bolaños, D., Cole, R. A., Ward, W., Borts, E., and Svirsky, E. (2011). Flora: Fluent oral reading assessment of children's speech. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(4):1–19.

Bolanos, D., Cole, R. A., Ward, W. H., Tindal, G. A., Schwanenflugel, P. J., and Kuhn, M. R. (2013). Automatic assessment of expressive oral reading. *Speech Communication*, 55(2):221–236.

Carchedi, L. C., Barrére, E., and de Souza, J. F. (2021). Avalia online: um sistema para avaliação em larga escala de testes de fluência de leitura. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 01–11. SBC.

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.

Cheng, J., Chen, X., and Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. *Speech Communication*, 73:14–27.

Duchateau, J., Cleuren, L., Ghesquière, P., et al. (2007). Automatic assessment of children's reading level. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1210–1213.

Fan, R., Afshan, A., and Alwan, A. (2021). Bi-apc: Bidirectional autoregressive predictive coding for unsupervised pre-training and its application to children's asr. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7023–7027. IEEE.

Hagen, A. and Pellom, B. (2005). A multi-layered lexical-tree based recognition of subword speech units. *Proc. L&TC, Poznan, Poland*.

Hagen, A., Pellom, B., and Cole, R. (2007). Highly accurate children's speech recognition for interactive reading tutors using subword units. *speech communication*, 49(12):861–873.

Hagen, A., Pellom, B., Van Vuuren, S., and Cole, R. (2004). Advances in children's speech recognition within an interactive literacy tutor. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 25–28.

Jain, R., Yiwere, M., Bigioi, D., and Corcoran, P. (2022). Can self-supervised learning solve the problem of child speech recognition? *arXiv preprint arXiv:2204.05419*.

Junior, A. C., Casanova, E., Soares, A., de Oliveira, F. S., Oliveira, L., Junior, R. C. F., da Silva, D. P. P., Fayet, F. G., Carlotto, B. B., Gris, L. R. S., et al. (2021). Coraa: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *arXiv preprint arXiv:2110.15731*.

Metallinou, A. and Cheng, J. (2014). Using deep neural networks to improve proficiency assessment for children english language learners. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Sklar, M. B., and Tobin, B. (2003). Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 29(1):61–117.

Poulsen, R., Hastings, P., and Allbritton, D. (2007). Tutoring bilingual students with an automated reading tutor that listens. *Journal of Educational Computing Research*, 36(2):191–221.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

Reeder, K., Shapiro, J., and Wakefield, J. (2007). The effectiveness of speech recognition technology in promoting reading proficiency and attitudes for canadian immigrant children. In *15th European Conference on Reading*.

Ruiz, N. and Federico, M. (2015). Phonetically-oriented word error alignment for speech recognition error analysis in speech translation. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 296–302. IEEE.

Sabu, K. and Rao, P. (2018). Automatic assessment of children's oral reading using speech recognition and prosody modeling. *CSI Transactions on ICT*, 6(2):221–225.

Sabu, K., Swarup, P., Tulsiani, H., and Rao, P. (2017). Automatic assessment of children's l2 reading for accuracy and fluency. In *SLaTE*, pages 121–126.

Sorgatto, D. W., Nogueira, B. M., Cáceres, E. N., and Mongelli, H. (2021). Avaliação de classificadores para relacionar características escolares a indicadores educacionais. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 1232–1242. SBC.

Tao, J., Ghaffarzadegan, S., Chen, L., and Zechner, K. (2016). Exploring deep learning architectures for automatically grading non-native spontaneous speech. In *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 6140–6144. IEEE.

Tchistiakova, S. (2019). Time delay neural network. `https://kaleidoescape.github.io/tdnn/`.

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

Yilmaz, E., Pelemans, J., et al. (2014). Automatic assessment of children's reading with the flavor decoding using a phone confusion model. *Proceedings Interspeech 2014*, pages 969–972.

Yu, F., Yao, Z., Wang, X., An, K., Xie, L., Ou, Z., Liu, B., Li, X., and Miao, G. (2021). The slt 2021 children speech recognition challenge: Open datasets, rules and baselines. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1117–1123. IEEE.

Zechner, K., Sabatini, J., and Chen, L. (2009). Automatic scoring of children's read-aloud text passages and word lists. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18.