CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Identificação de silabação em áudios de leitura de crianças em anos iniciais

Bruno de O. Jucá[2], Caio C. Rocha[2], Rômulo C. de Mello[2], Eduardo Barrére[1,2], Jairo Francisco de Souza[1,2]

[1]LApIC Research Group – Universidade Federal de Juiz de Fora

[2]Departamento de Ciência da Computação – Universidade Federal de Juiz de Fora

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

Abstract. *This paper focuses on automatically detecting syllabification in the speech of children during the literacy phase, which poses a challenge in the assessment of reading fluency. Automatic Speech Recognition (ASR) allows fast processing of speech recordings, generating acoustic metrics about syllable duration and the duration of pauses between syllables. Therefore, we propose heuristics to automatically classify syllabification in speech in a straightforward way using those features. The proposal achieved an accuracy of 0.87 on the validation set, highlighting that the automatic classification of syllabification in speech can be applied in fluency assessments.*

Resumo. *Este artigo aborda a detecção automática de silabação em áudios de crianças em fase de alfabetização, que é um desafio da avaliação de fluência em leitura. Nesse contexto, o reconhecimento automático da fala (ASR) permite processar os áudios de forma rápida e objetiva, gerando métricas acústicas sobre a duração das sílabas e a duração dos intervalos entre elas. Assim, propõe-se neste trabalho a aplicação de heurísticas que usam essas características para classificar automaticamente a silabação. Os resultados obtidos alcançaram acurácia de 0,87 em uma base de validação, o que destaca que a classificação automática da silabação pode ser aplicada na avaliação de fluência.*

1. Introdução

A avaliação da fluência em leitura no Ensino Fundamental é um tema relevante na área de avaliação da educação, pois a fluência em leitura é indicativa da compreensão textual [Martins and Capellini 2019]. As avaliações por meio de provas de leitura oral são importantes para determinar o desempenho na leitura e permitem caracterizar os tipos de erros de modo a perceber as estratégias utilizadas no processo de aprendizagem [Puliezi and Maluf 2014]. Dado o desenvolvimento do processamento de linguagem natural (PLN), o reconhecimento automático de fala (*Automatic Speech Recognition* – ASR) é um aliado promissor das avaliações de fluência de leitura em larga escala [Soares et al. 2018], que possuem milhões de dados para serem avaliados.

Primeiramente, as avaliações em larga escala podem fornecer dados concretos e precisos sobre o desempenho geral dos estudantes no aprendizado, permitindo a identificação de tendências e de lacunas no ensino [Carchedi et al. 2021]. À vista disso,

pode-se citar as avaliações conduzidas pela Associação Bem Comum (ABC), como a Parceria pela Alfabetização em Regime de Colaboração (PARC)[1]. Pode-se citar também as avaliações ofertadas pelo Governo Federal, como a Prova Brasil, a Avaliação Nacional da Alfabetização (Ana) e a Avaliação Nacional da Educação Básica (Aneb), que compõem o Sistema Nacional de Avaliação da Educação Básica (Saeb). Essas avaliações permitem acompanhar o nível de aprendizado de estudantes em escolas públicas [de Castro and Callou 2022].

Entretanto, tais avaliações dependem da disponibilidade de profissionais qualificados para ir às escolas, coletar os áudios das leituras dos alunos, escutar os áudios e avaliá-los manualmente, o que demanda tempo e alto aporte financeiro. Em vista disso, o desenvolvimento de tecnologias de avaliação automática permite reduzir os recursos humanos e financeiros necessários para realizar esses tipos de avaliação [Carchedi et al. 2021]. Outrossim, uma solução que utiliza tecnologias de processamento do sinal de fala humana e métricas de avaliação possibilita a análise objetiva da fluência, que é muito importante para uma quantificação exata do nível de ensino [Almeida Silva et al. 2021].

Desse ponto de vista, os modelos de ASR podem agregar agilidade e precisão à avaliação, pois podem gerar a transcrição automática de áudios e podem gerar informações detalhadas sobre o sinal de fala. Os modelos Transformers, como o Wav2Vec2 [Baevski et al. 2020], produzem bons resultados para tarefas gerais de transcrição e são adaptáveis para tarefas específicas [Evrard 2023], como no exemplo da transcrição de áudios de leituras infantis, caso sejam refinados para isso.

Dentre as métricas de fluência que podem ser calculadas, a detecção de silabação, ou seja, a pronúncia indesejadamente pausada ou prolongada de sílabas das palavras, é um desafio, pois este é um fator subjetivo da avaliação. A título de exemplo, um aluno que possui um ritmo de leitura lento terá pausas mais longas entre as sílabas das palavras se comparado a outro que possui um ritmo mais rápido. Outrossim, demais fatores que influenciam na duração das sílabas na fala são o estresse dado a cada grupo silábico e o número de fonemas por sílaba [Crystal and House 1990]. Por conseguinte, é difícil definir uma métrica universal que capture a silabação.

Contudo, a identificação da silabação é importante para a avaliação de fluência, pois ela está ligada ao ritmo de leitura e à compreensão do que foi lido. Um estudo destaca, a partir da análise dos dados de leitura do *National Assessement of Educational Progress* (NAEP), programa realizado nos EUA, que a leitura oral tem correlação positiva com a compreensão da leitura [Puliezi and Maluf 2014]. Por outro lado, a identificação automática de silabação não é uma tarefa trivial, visto que não existem parâmetros bem definidos do que é uma leitura silabada, cabendo a decisão de um avaliador experiente para cada leitura em avaliações manuais. Ainda, leituras de crianças em fase de alfabetização são naturalmente mais pausadas, com muita presença de falsos começos, erros de leitura e hesitações, o que dificulta ainda mais a sua avaliação.

Este artigo explora o problema de identificação de silabação em áudios de crianças em fase de alfabetização, que é um problema de classificação binária, com o objetivo de auxiliar sistemas de avaliação automática de fala que têm sido utilizados em avaliação educacional. Para isso, são apresentadas heurísticas que usam características acústicas

[1] `https://abemcomum.org/parceria-pela-alfabetizacao-em-regime-de-colaboracao/`

para classificar essas leituras, levando em conta as sílabas pronunciadas e o intervalo entre essas sílabas. As heurísticas foram avaliadas em um conjunto de áudios de leituras de crianças dos anos letivos iniciais do ensino fundamental. Os experimentos mostram uma boa acurácia e uma precisão alta da solução apresentada. Por fim, até onde os autores possuem conhecimento, este é o primeiro trabalho no contexto brasileiro a analisar soluções para identificação de silabação durante o processo de alfabetização na língua portuguesa.

2. Trabalhos relacionados

Não foram encontrados artigos que busquem identificar especificamente a silabação, analisando as pausas entre sílabas e o prolongamento de sílabas em leituras. Apesar disso, são apresentados nesta seção trabalhos sobre a correção de testes de leitura, sobre a silabificação automática de palavras e sobre a segmentação silábica de pronúncias em áudios, ou seja, a identificação dos limites de separação das sílabas no áudio. Esses trabalhos se aproximam, em certa medida, ao que é proposto neste artigo já que apresentam técnicas diferentes para identificar os limites silábicos, algo necessário para, posteriormente, quantificar os intervalos e definir silabações no contexto de leituras de crianças em processo de alfabetização.

[de Assis et al. 2022] apresentam um método para correção automática da leitura de pseudopalavras a partir de áudios de crianças em fase de alfabetização. Com a transcrição automática gerada por um modelo de ASR, são aplicadas heurísticas para determinar se as pseudopalavras foram lidas corretamente. O trabalho aponta que silabações são um desafio no processo de avaliação, porém a avaliação final sobre a leitura de cada palavra é binária (correta ou incorreta), sem que haja uma identificação da ocorrência de silabação.

[Boháč et al. 2016] propõem um esquema para silabificar automaticamente palavras reconhecidas de áudios em tcheco e identificar, por meio do alinhamento forçado, as sílabas e seus fonemas com o intervalo de tempo associado. Contudo, a abordagem adotada realiza uma silabificação das palavras transcritas antes de fazer um alinhamento das palavras silabificadas com uma transcrição fonética da fala. A última está associada à informação de tempo de ocorrência de cada fonema. Dessa forma, a identificação das sílabas no áudio segue as regras gramaticais de separação silábica.

[Panda and Nayak 2016] apresentam uma técnica de segmentação automática de sinais de fala baseada em sílabas em várias línguas indianas. Segundo o artigo, a representação da fala no domínio do tempo é extraída do áudio e é utilizada para identificar as vogais presentes. Então, as sílabas são identificadas e o sinal de fala é segmentado por unidade silábica, produzindo uma representação da fala segmentada por sílabas. Com isso, a segmentação é realizada com base na representação das sílabas por grupos de vogais, independentemente da língua em questão.

[Karim and Suyanto 2019] propõem um modelo de segmentação silábica automática baseado em características da energia ao longo do tempo que utiliza um limiar otimizado por meio de algoritmos genéticos (GA) para encontrar as fronteiras de cada sílaba. Para isso, ele utiliza uma abordagem iterativa para dividir as palavras e para aglutinar sílabas que tenham um núcleo silábico comum. O benefício é que a identificação silábica possui um limiar adaptativo, que mostrou maior acurácia em comparação a um modelo com um limiar fixo de separação silábica. Entretanto, o estudo foi realizado com 110 leituras em indonésio gravadas por somente uma falante, o que põe em questão se ele pode

ser generalizado para outros falantes.

[N et al. 2020] sugerem um sistema de detecção e de correção automática de gagueira que utiliza *Mel Frequency Cepstral Coefficients* (MFCC) e *Linear Predictive Coefficients* (LPC) para extrair as características do áudio. As principais características utilizadas são a energia a curto prazo e a correlação entre trechos do áudio. Dessa forma, é possível identificar repetições e prolongações de sílabas que ocorrem na fala. Mesmo que esse sistema possa ajudar na identificação da silabação do ponto de vista do prolongamento de sílabas, ele é voltado especificamente para a identificação da gagueira em amostras de áudios de pessoas que possam apresentar esse distúrbio.

Apesar da relevância desses trabalhos para encontrar soluções de silabificação de palavras, isso não é suficiente para classificar a silabação, que é um aspecto importante da avaliação de fluência, pois somente as crianças que têm leituras não silabadas podem ser consideradas fluentes. Portanto, o presente artigo contribui para a avaliação qualitativa da fluência em leitura, pois apresenta uma solução de uma das características da fala fluente ainda pouco explorada na avaliação de leitura oral em português.

3. Materiais e métodos

Dada a motivação do trabalho, desenvolveram-se estratégias para classificar as silabações, as quais são apresentadas a seguir. Em suma, o algoritmo proposto utiliza uma rede neural de ASR para transcrever os áudios que contêm as pronúncias das palavras e um alinhador forçado que associa as transcrições às informações de início e fim de cada sílaba de cada palavra, o que permite analisar a duração das sílabas e a confiança atribuída pelo modelo a cada pronúncia. Por fim, foram desenvolvidas heurísticas que utilizam essas informações para gerar as classificações de silabação. A Figura 1 resume as etapas realizadas, que serão detalhadas ao longo dessa seção.

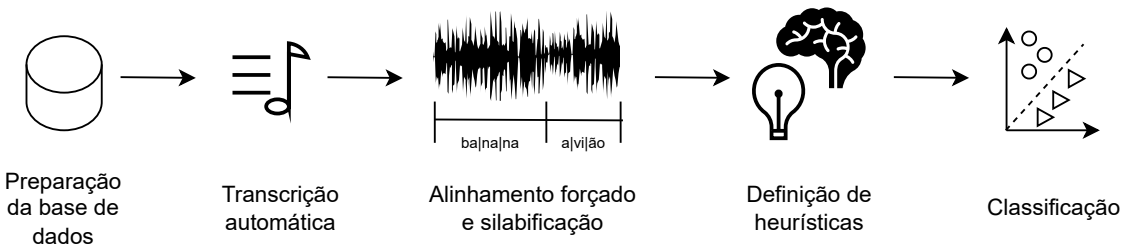

Figura 1. Etapas realizadas para identificação de silabações

3.1. Preparação da base de dados

O conjunto de leituras utilizado neste estudo pertence à avaliação de fluência da avaliação PARC 2021, na qual foram aplicados testes de leitura para crianças do Ensino Fundamental de escolas do estado do Maranhão. Os materiais utilizados para fazer os testes de leitura são 5 listas de 60 palavras cada, as quais são lidas pelas crianças e gravadas pelo(a) professor(a). Com o objetivo de definir uma base de dados sobre silabação, os autores selecionaram e avaliaram leituras de palavras individuais dos testes. As anotações foram feitas por 3 avaliadores que dividiram a quantidade de áudios a serem avaliados por cada um. O processo de anotação dos áudios foi realizado manualmente, ou seja, os

avaliadores escutavam os áudios e marcavam numa planilha as informações de início e de fim da leitura de cada palavra, se houve silabação, e os índices dos intervalos entre sílabas onde ocorreu essa silabação, entendida como uma pausa perceptível na leitura da palavra. A Tabela 1 exemplifica essas anotações para um áudio.

ID do áudio	Palavra	Início	Fim	Índice do intervalo em que ocorreu silabação	Silabou
01	banana	16,50	17,88	1;2	Sim
01	avião	18,70	20,56	2	Sim
01	bolo	21,74	23,22	-	Não
01	panela	25,32	26,9	1	Sim

Tabela 1. Exemplo de anotações realizadas manualmente sobre um áudio

Foram selecionadas 303 amostras de leitura, das quais 143 são silabadas e 160 são não silabadas, mantendo o equilíbrio entre os dados. As leituras foram feitas por 57 alunos e, além disso, todas as leituras apresentam a pronúncia correta das palavras. A partir dessas anotações, foram montados dois *datasets* para o escopo deste estudo: um *dataset* de testes e um *dataset* de validação. O *dataset* de testes foi utilizado na definição das heurísticas e consiste em 212 leituras, sendo 100 silabadas e 112 não silabadas. Já o *dataset* de validação foi utilizado para avaliar as heurísticas propostas e possui 91 leituras, sendo 43 leituras silabadas e 48 não silabadas.

Em seguida, os áudios foram transcritos usando um modelo de rede neural Wav2Vec2 XLSR-53 refinado para o português. O dataset utilizado no refinamento, chamado CORAA (Corpus of Annotated Audios) v1, é composto por 290,77 horas de áudios em português brasileiro. O refinamento do modelo, realizado por [Junior et al. 2021], se deu com a GPU NVIDIA TESLA V100 32GB, usando um tamanho de batch de 8 e acumulação de gradiente sobre 24 etapas.

Para auxiliar na transcrição, construiu-se um modelo de língua, que é um arquivo auxiliar que permite o ajuste de pequenas variações na transcrição segundo um dicionário de possíveis combinações de palavras. Essas combinações foram representadas por unigramas e por bigramas que correspondem às palavras individuais e a todos os pares possíveis de palavras contidas na lista de palavras, respectivamente. Assim, a transcrição se aproxima da lista de palavras esperadas, o que ajuda na identificação das palavras lidas. Com uma melhor identificação das palavras lidas, é possível definir com maior grau de certeza os intervalos de pronúncia das sílabas das palavras. Vale ressaltar que o modelo de língua não torna a rede enviesada, pois altera somente a transcrição do áudio e não influencia o reconhecimento das palavras.

3.2. Identificação dos limites entre as sílabas

Para silabificar as palavras, foram utilizadas duas formas de separação silábica: a separação canônica e a separação fonética, a fim de comparar qual seria a melhor no escopo do artigo. A separação canônica tem como objetivo dividir as palavras de acordo com as regras gramaticais, como em $sor\text{-}ri\text{-}so$. Por outro lado, na separação fonética, as palavras são divididas em sílabas de acordo com a forma como os sons são produzidos na fala, como em $so\text{-}rri\text{-}so$. Para isso, é necessário ter conhecimento dos elementos

fonéticos da língua portuguesa, como hiatos, ditongos, dígrafos, encontros consonantais, entre outros. Contudo, segundo os testes realizados, não se identificou nenhuma diferença prática entre as duas formas de silabificação e, então, decidiu-se por adotar a silabificação canônica, que é mais simples.

Como próxima etapa, foi realizado um alinhamento de sequências entre as transcrições geradas pelo modelo e as palavras das listas de palavras correspondentes. Esse alinhamento compara as duas sequências de entrada e produz um relatório de palavras corretas (C), substituições (S), inserções (I) e deleções (D) da transcrição em relação à sequência de palavras de referência, conforme apresentado na Tabela 2.

Referência	banana	avião	bolo	
Transcrição	banana		bola	pá
Alinhamento	C	D	S	I

Tabela 2. Alinhamento de sequências

Para os próximos experimentos, apenas as palavras corretas e as substituições foram consideradas, pois deleções e inserções não estão relacionadas com a identificação de silabação.

Em seguida, o alinhamento forçado foi utilizado para encontrar as fronteiras temporais das leituras de cada palavra, a partir das informações de alinhamento entre a transcrição automática e as palavras de referência, segundo a Figura 2.

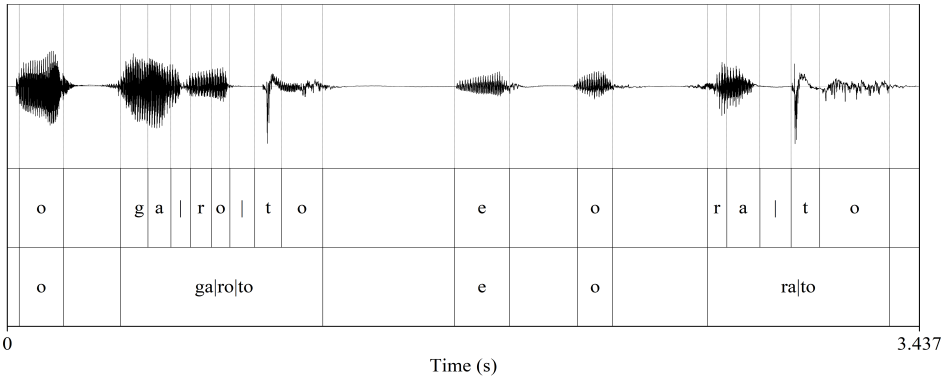

Figura 2. Exemplo de alinhamento forçado

O alinhamento forçado é capaz de produzir métricas relevantes para a avaliação de fluência [Gomes Jr et al. 2019], gerando as marcações temporais de cada palavra e a probabilidade associada a sua pronúncia. Essa técnica possui 3 etapas principais, que são estimar a probabilidade de cada letra por *frame* da representação vetorial do áudio; gerar a matriz da treliça, que representa a probabilidade das letras por instante de tempo; e encontrar o caminho ótimo da matriz da treliça para cada palavra de referência.

3.3. Definição das heurísticas de classificação

As estratégias utilizadas para definir as heurísticas baseiam-se em utilizar combinações entre duração da sílaba anterior (DS), duração do intervalo intersilábico (DI),

ambos em milissegundos, e probabilidade do intervalo intersilábico (PI), definindo limiares de discriminação. A título de comparação com as estratégias que serão definidas, a Figura 3a representa uma leitura não silabada. Visto isso, foram propostas 3 estratégias para definir as classificações, de acordo com as Equações 1, 2 e 3.

$$(PI > PI_{limiar} \ \wedge \ DI > DI_{limiar}) \tag{1}$$

$$(DS > DS_{limiar}) \tag{2}$$

$$(PI > PI_{limiar} \ \wedge \ DI > DI_{limiar}) \ \vee \ (DS > DS_{limiar}) \tag{3}$$

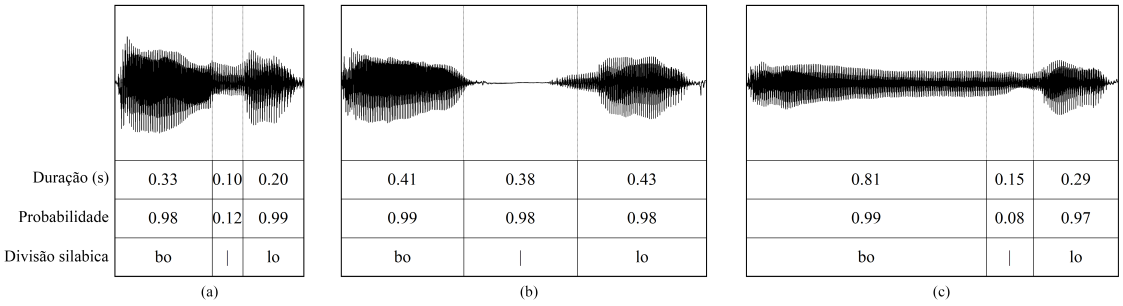

Figura 3. Exemplo de diferentes pronúncias da palavra "bolo".

A estratégia definida na Equação 1 consiste em analisar se a probabilidade do intervalo silábico (PI) e a duração desse intervalo (DI) são maiores que dois limiares previamente definidos. Nesse sentido, parte-se do princípio de que há uma chance considerável de ter ocorrido, de fato, um intervalo, e se esse intervalo foi prolongado, deve haver uma silabação. Um exemplo de ocorrência de uma leitura silabada que se busca classificar com base nessa estratégia está ilustrada na Figura 3b.

De acordo com a estratégia definida na Equação 2, se a duração da sílaba anterior (DS) for maior que um limiar, ou seja, se for uma pronúncia prolongada, deve ter ocorrido uma silabação, já que a prolongação da sílaba também é uma característica das leituras silabadas ($DS > DS_{limiar}$). A Figura 3c ilustra o tipo de leitura que se busca classificar com base nessa estratégia.

Por fim, a estratégia definida na Equação 3 leva em consideração as duas estratégias definidas pelas Equações 1 e 2 em conjunto.

Inicialmente, para buscar valores que servissem como um ponto de partida para esses limiares, foram levantados alguns dados estatísticos para analisar o comportamento dessas métricas na base, de acordo com a Tabela 3.

Assim, foi possível dimensionar a distribuição dos valores escolhidos para servirem como características de classificação de silabação de acordo com as estratégias

	PI	***DI (s)***	***DS (s)***
Mínimo	0	0,02	0,02
Máxima	0,998	6,88	1,48
Mediana	-	0,53	0,38
Média	0,49	0,83	0,42
Desvio Padrão	0,38	0,95	0,29

Tabela 3. Valores estatísticos dos dados relativos as sílabas e as pausas entre sílabas

definidas. Como são infinitas as combinações possíveis para valores dos limiares a serem definidos, decidiu-se por realizar uma busca exaustiva entre as combinações possíveis, dentro do universo de possíveis valores, com o intuito de definir heurísticas que consigam classificar as leituras entre silabadas e não silabadas de maneira precisa. Os intervalos considerados na busca foram $0,00 \leq PI_{limiar} \leq 1,00$, com passo de $0,01$, $0,0 \leq DI_{limiar} \leq 6,9$, com passo de $0,1$, e $0,0 \leq DS_{limiar} \leq 1,5$, com passo de $0,1$.

4. Resultados

A busca exaustiva realizada para encontrar os melhores limiares para a definição das heurísticas retornou, para cada uma das estratégias definidas anteriormente, as heurísticas presentes na Tabela 4. Foram selecionados os valores que obtiveram melhores acurácias para cada estratégia.

Heurística	Combinações
H1	$PI \geq 0,01 \wedge DI > 0,2$
H2	$DS > 0,3$
H3	$(PI \geq 0,01 \wedge DI > 0,2) \vee DS > 0,5$

Tabela 4. Heurísticas utilizadas, onde pelo menos uma sílaba da palavra deve satisfazer as condições.

O gráfico da Figura 4 mostra o resultado da busca exaustiva realizada para encontrar os valores da H3. Percebe-se que os maiores valores de acurácia ocorrem quando PI_{limiar} se aproxima de zero. Esse comportamento foi o mesmo na busca realizada para encontrar H1, embora fosse esperado que esse limiar de decisão tivesse um valor maior.

Com as heurísticas definidas, foram encontrados os valores descritos na Tabela 5.

Diante das métricas apresentadas, a H3 conseguiu separar mais claramente as leituras silabadas das não silabadas, como indica o coeficiente de correlação de Matthews (MCC). De fato, sua precisão alta para a classe de silabados indica que ela identificou corretamente muitas leituras classificadas como silabadas. Para a classe de não silabados, a precisão também é alta, apesar de ser menor que da classe de silabados. Aliás, a H3 selecionou a maioria das leituras não silabadas, como indica a sua revocação alta. Porém, para a classe de silabados, a revocação obtida é consideravelmente menor, o que aponta que há uma parcela das leituras silabadas que não foram identificadas.

A heurística H1 apresentou maior precisão para a classe de silabados e maior revocação para a classe de não silabados, pois classificou a maioria das leituras como não

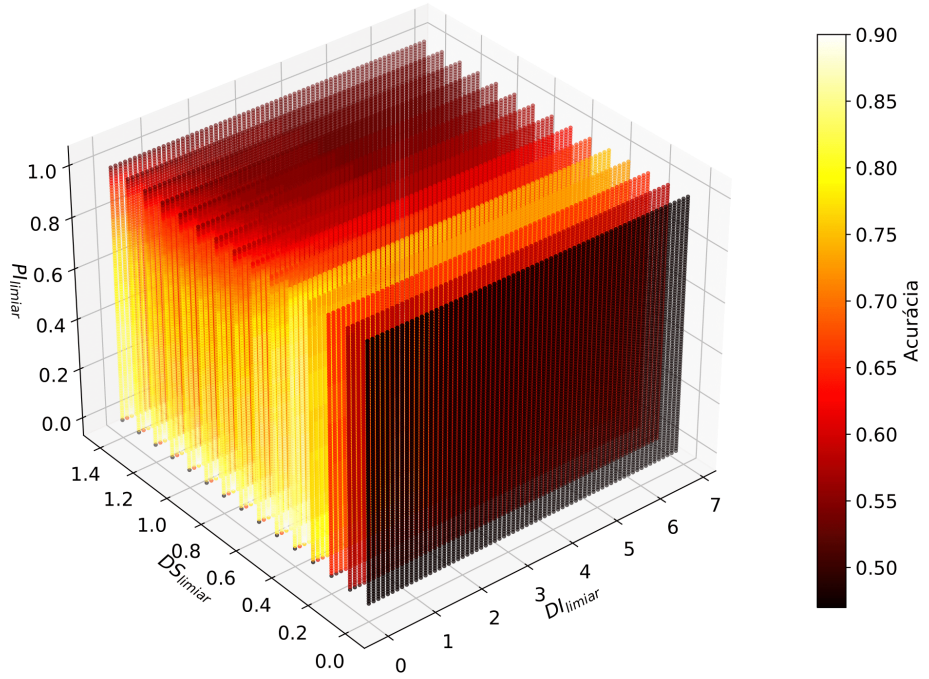

Figura 4. Gráfico das combinações de valores da estratégia 3 obtidos pela busca exaustiva

Dataset	Heurística	Acurácia	Precisão (positiva)	Revocação (positiva)	Precisão (negativa)	Revocação (negativa)	MCC
Teste	H1	0,89	**0,97**	0,78	0,83	**0,98**	0,78
	H2	0,73	0,70	0,74	0,75	0,71	0,45
	H3	**0,90**	0,94	**0,84**	**0,87**	0,96	**0,80**
Validação	H1	0,84	**0,94**	0,70	0,78	**0,96**	0,69
	H2	0,70	0,65	**0,79**	0,77	0,62	0,42
	H3	**0,87**	0,92	**0,79**	**0,83**	0,94	**0,74**

Tabela 5. Métricas das melhores heurísticas obtidas com a busca exaustiva, onde "positiva" e "negativa" denotam as classes de silabados e não silabados, respectivamente.

silabadas. Assim, ela obteve uma revocação menor para a classe de silabados, pois muitas leituras silabadas foram classificadas como não silabadas. De forma semelhante, sua precisão menor para a classe de não silabados indica que muitas das leituras classificadas como não silabadas eram silabados que não foram identificados.

Ao contrário da H1, a H2 apresentou maior precisão para a classe de não silabados e maior revocação para a classe de silabados, pois classificou a maioria das leituras como silabadas. Desse ponto de vista, a H3 alcançou um equilíbrio entre a H1 e a H2, apresentando os melhores resultados.

Por fim, percebe-se que todas as heurísticas obtiveram resultados melhores na base de teste, pois ela foi utilizada na própria escolha dos limiares de classificação, porém os resultados da validação repetem os padrões constatados na base de teste e confirmam a eficácia das heurísticas.

5. Considerações finais

Foram apresentadas três heurísticas para classificação de silabação em áudios de leituras de crianças em fase de alfabetização. Observou-se que algumas das características se mostraram mais significativas do que outras, como a duração da sílaba anterior *versus* a probabilidade do silêncio entre as sílabas, diferentemente do que se esperava inicialmente.

Após uma análise dos dados brutos, notou-se que a utilização da marcação de silêncio, representada pelo símbolo "|", no processo de alinhamento forçado, tende a funcionar como um ponto de espera entre as sílabas. Isso acontece porque a duração do silêncio se estende enquanto a probabilidade da primeira letra da sílaba subsequente é baixa. Em outras palavras, durante o alinhamento forçado do áudio, o algoritmo busca encontrar um conjunto de *frames* de áudio que represente o símbolo "|". Mesmo quando o modelo tem incertezas na atribuição de probabilidade ao silêncio (geralmente atribuindo uma probabilidade inferior a 5%), essa probabilidade ainda é maior do que a probabilidade da próxima letra, o que leva à prolongação da duração do "|". Isso fornece informação suficiente para a tomada de decisão quanto à silabação. Várias situações podem levar a uma baixa probabilidade para a marcação de silêncio, como uma sílaba sendo pronunciada de forma prolongada (o que é comum em hesitações na leitura de crianças) ou a presença de ruídos no áudio (uma característica comum nos áudios desta base de dados, devido à baixa qualidade dos equipamentos e falta de isolamento acústico no ambiente). Esses cenários devem ser explorados mais detalhadamente em pesquisas futuras. No entanto, é importante ressaltar que os resultados obtidos são promissores e indicam que esse método pode ser aplicado com sucesso na avaliação da fluência na leitura, mesmo em áudios com essas características.

É importante considerar que este estudo apresenta limitações. O alinhamento forçado pode apresentar erros, o que resulta em imprecisões nas identificações dos limites silábicos reais. Quanto mais precisa for a técnica, diante de um bom modelo acústico, mais precisa será a identificação dos limites silábicos e das características de probabilidade coletadas. Outro aspecto a se considerar diz respeito ao grau de subjetividade presente na avaliação manual e a possíveis erros nas classificações humanas.

Neste trabalho não foi realizada a identificação do ritmo de leitura de cada estudante, mas esse, possivelmente, seria um fator útil a se considerar. Pesquisas anteriores [Puliezi and Maluf 2014, Miller and Schwanenflugel 2008] mostram que o ritmo de leitura está ligado à fluência oral, portanto sua utilização na classificação de leituras silabadas poderia gerar resultados mais acurados e mais precisos. Além disso, outras características de prosódia podem adicionar informações pertinentes para essa classificação, pois a definição de fronteiras entre as sílabas das palavras é uma característica subjetiva influenciada por diversos fatores [Crystal and House 1990].

Vale destacar que o fato de uma leitura apresentar ou não silabações não é um fator decisivo para apontar a fluência em leitura, mas quando considerado em conjunto com outros aspectos da leitura, é possível separar os leitores em diferentes categorias que buscam descrever o panorama da alfabetização em um grupo de crianças. Pesquisas envolvendo avaliação automática de leitura são importantes no desenvolvimento de recursos que auxiliem a identificar o andamento da alfabetização de uma população. No Brasil, há poucas pesquisas em avaliação automática de fluência e torna-se relevante estimular a apresentação de ideias que promovam o avanço do conhecimento na área.

Referências

Almeida Silva, W., Carchedi, L., Gomes Jr, J., Souza, J., Barrere, E., and Souza, J. (2021). A framework for large-scale automatic fluency assessment. *International Journal of Distance Education Technologies*, 19.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Boháč, M., Matějů, L., Rott, M., and Šafařík, R. (2016). Automatic syllabification and syllable timing of automatically recognized speech – for czech. In Sojka, P., Horák, A., Kopeček, I., and Pala, K., editors, *Text, Speech, and Dialogue*, pages 540–547, Cham. Springer International Publishing.

Carchedi, L., Barrére, E., and de Souza, J. (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, Porto Alegre, RS, Brasil. SBC.

Crystal, T. H. and House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America*, 88 1:101–12.

de Assis, E., Ferreira, A. L., Silva, C., and de Souza, J. (2022). Classificação automática de áudios de leituras de pseudopalavras para avaliação em larga escala de fluência da leitura de crianças em fase de alfabetização. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 27–38, Porto Alegre, RS, Brasil. SBC.

de Castro, M. H. G. and Callou, R. (2022). *Educação em Pauta 2022: Desafios da Educação Básica no Brasil*. Organização dos Estados Ibero-americanos.

Evrard, M. (2023). *Transformers in Automatic Speech Recognition*, pages 123–139. Springer International Publishing.

Gomes Jr, J., Almeida Silva, W., Souza, J., Barrere, E., and Souza, J. (2019). Uso de alinhadores forçados para avaliação automática em larga escala da fluência em leitura. In *Anais do XXX Simpósio Brasileiro de Informática na Educação*, page 61, Porto Alegre, RS, Brasil. SBC.

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., and Aluísio, S. M. (2021). Coraa: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese.

Karim, R. and Suyanto, S. (2019). Optimizing parameters of automatic speech segmentation into syllable units. *International Journal of Intelligent Systems and Applications*, 11:9–17.

Martins, M. A. and Capellini, S. A. (2019). Relação entre fluência de leitura oral e compreensão de leitura. *CoDAS*, 31(1):e20170244.

Miller, J. and Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly*, 43(4):336–354.

N, A., S, K., D, K., Chanda, P., and Tripathi, S. (2020). Automatic correction of stutter in disfluent speech. *Procedia Computer Science*, 171:1363–1370.

Panda, S. P. and Nayak, A. K. (2016). Automatic speech segmentation in syllable centric speech recognition system. *International Journal of Speech Technology*, 19:9–18.

Puliezi, S. and Maluf, M. (2014). A fluência e sua importância para a compreensão da leitura. *Psico-USF*, 19:467–475.

Soares, E., Carchedi, L., Gomes Jr, J., Barrere, E., and Souza, J. (2018). Avaliação automática da fluência em leitura para crianças em fase de alfabetização. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, page 11, Porto Alegre, RS, Brasil. SBC.