Finite-State Transducers for Oral Spelling Detection

Gabriel J. R. Soares[1], **José E. C. Silva**[1], **Jairo F. de Souza**[1,2]

[1]LApIC Research Group - Universidade Federal de Juiz de Fora (UFJF)

[2]Department of Computer Science – Universidade Federal
Juiz de Fora – MG – Brazil

{gabriel.soares, jose.carvalho}@fundacaocaed.org.br, jairo.souza@ufjf.br

Abstract. *Reading fluency assessment plays a central role in early education systems worldwide. Countries such as the United States and Brazil administer large-scale oral reading assessments to monitor educational outcomes and guide intervention. However, most of the automatic assessments are often coarse in granularity. As a result, they are poorly equipped to handle children who do not yet decode words fluently and instead rely on spelling out individual letters or syllables. We show that finite-state transducers can be used to detect spelling to improve oral reading assessments. We demonstrate the effectiveness of our method on a corpus of annotated child speech, showing that it provides insight into early decoding strategies.*

1. Introduction

Reading fluency assessment plays a central role in K-12 education systems worldwide, with countries such as the United States and Brazil administering large-scale oral reading assessments to monitor educational outcomes and guide intervention [Silva et al. 2022]. These assessments are vital for understanding broad trends and identifying schools or regions needing support. However, a significant challenge arises when evaluating children who do not yet decode words fluently.

Decoding, the ability to map printed letters to their corresponding sounds, is the foundational skill upon which fluent reading is built [Gough and Tunmer 1986]. Foundational to this process are letter-sound knowledge and phonemic awareness, i.e., the understanding of the relationship between printed letters and their sounds. Research indicates that early proficiency in these skills is a strong predictor of later reading success and overall academic achievement, whereas children who rely on non-fluent decoding strategies are at greater risk for long-term reading difficulties [Hulme and Snowling 2013]. These challenges can subsequently lead to reduced reading volume, limited vocabulary growth, and decreased motivation [Stanovich 1986].

Detecting when a child spells out letters accurately is crucial for early diagnosis. Such behavior indicates that the child has acquired some letter-sound knowledge but struggles to blend those units into fluid word recognition. While deficits in letter-sound knowledge are causally linked to reading problems, they are also remediable with appropriate instruction [Hulme and Snowling 2013]. Assessments capable of capturing letter-by-letter or syllable-level decoding provide a window into a child's developing skills, allowing for specific interventions that strengthen foundational competencies and promote

DOI: 10.5753/sbie.2025.12213

fluent reading. In contrast, coarse-grained scoring that simply marks any spelled or syllabified word as incorrect misses this valuable diagnostic information and delays targeted support.

Prior work in [Rocha et al. 2024] specifies three fluency profiles: fluent, pre-reader and beginner. A child's profile is determined based on the number of words read, and the fraction of those that were read correctly. The pre-reader profile is further divided into levels based on the number of words a child has syllabified correctly, and another based on the number of words a child did spelled correctly. However, most research to date has employed a binary "correct/incorrect" reading, and hence cannot assess in which specific sub-level a pre-reader child is in.

In the Brazilian Portuguese setting, this problem is further complicated by regional variation in how letters are named. In particular, in the Northeast of Brazil, people often use an alternative naming convention for several letters, known informally as the *northeastern alphabet* [Luiz Gonzaga and Zé Dantas 1987]. Unlike regional accents that modify the pronunciation of phonemes, this variation involves entirely different phonemic realizations for letter names.

The objective of this study is to improve the detection of oral spelling in aloud reading audios from first years students. To that, we show Weighted Finite-State Transducers (see Section 2) can be effectively used to detect spelling in audios of Brazilian Portuguese children. To the best of our knowledge, this is the first study to apply Weighted Finite-State Transducers (WFSTs) to this problem for non-fluent Brazilian Portuguese speakers, demonstrating how this technology can be integrated into large-scale fluency assessments to provide the granular feedback needed for second-year elementary school students.

2. Background

In this work we rely on two main ideas: WFSTs and Wav2Vec2.

A WFST is an automaton that reads an input sequence, generates a corresponding output sequence, and accumulates an associated numerical weight. Formally, a WFST consists of a finite set of states and labeled transitions between them [Mohri et al. 2008]. Each transition specifies (i) an input symbol to be consumed, (ii) an output symbol to be produced, and (iii) a weight, which may represent a probability, a penalty or any other cost metric. As the transducer traverses a path of transitions, it produces an output sequence and sums the weights along that path, thereby yielding a total cost for the mapping from input to output.

Because they define a set of valid transductions (mappings from input to output sequences), FSTs are often referred to as *grammars* [Mohri et al. 2008]. In this sense, the FST acts as a formal grammar that defines the rules of a language, where the "language" is the set of all possible input-output string pairs with their associated weights.

In the context of reading assessment, WFSTs provide an intuitive way to model different reading patterns. Each reading strategy can be represented as a distinct path through the transducer. By designing appropriate transitions and weights, we can capture the characteristic patterns of each reading mode. Additionally, WFSTs can naturally handle common phenomena in children's reading such as hesitations, repetitions,

and self-corrections [Kouzelis et al. 2023][Neubig et al. 2012][Guo et al. 2025] through the strategic use of self-loops and ε-transitions (ε-transitions allow the WFST to move between states without consuming any input symbol, effectively modeling silent pauses or corrections while still accumulating the appropriate weight). See Figure 1 for an example of one of our WFSTs.

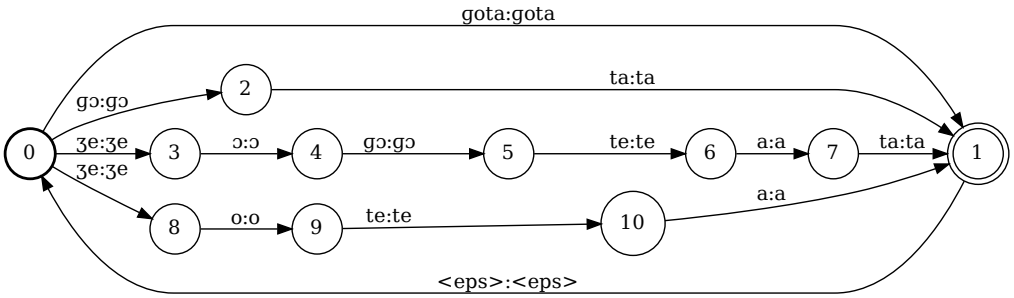

Figure 1. Word *gota* kernel in grammar G0. In a given transition (e.g. gɔ:gɔ), the first part corresponds to the symbol consumed from the input sequence (the audio), and the second to the symbol outputted by the WFST.

Wav2Vec2 is a self-supervised speech representation model that learns latent speech features from raw audio without requiring aligned transcriptions during pretraining [Baevski et al. 2020]. It consists of a convolutional encoder followed by a Transformer network trained to solve a contrastive prediction task. After this initial self-supervised pretraining phase, Wav2Vec2 can be fine-tuned for specific tasks with relatively small amounts of labeled data. For speech recognition tasks, the model is typically fine-tuned using a Connectionist Temporal Classification (CTC) approach [Graves et al. 2006], which aligns the model's predictions with text transcriptions without requiring precise frame-level alignments. The output of Wav2Vec2 for a given input audio is a sequence of frame-level probability distributions P over the vocabulary tokens (which, in our fine-tuned model, are phonemes). Here we refer to this output as the *emission matrix* of our acoustic model.

XSLR [Conneau et al. 2020] is a multilingual extension of Wav2Vec2 trained to learn cross-lingual speech representations from untranscribed audio in multiple languages. The model captures language-independent acoustic features that can be fine-tuned for downstream speech recognition tasks in individual languages, often yielding strong performance with limited labeled data. In our work, XSLR serves as the acoustic model that produces frame-level phoneme probabilities, which are then integrated with our WFST-based decoding framework to detect and classify different reading modalities, including spelling and syllabification patterns in children's oral reading.

3. Related Work

Most existing automatic reading-fluency systems focus on "correct versus incorrect" word-level errors or on high-fluency readers, without modeling the letter-by-letter spelling

strategies that pre-readers use. In addition, work in Brazilian Portuguese has not accounted for the northeastern alphabet. In what follows, we first summarize WFST-based approaches to miscue detection in children's reading and then review recent self-supervised acoustic models for child speech, highlighting how our method fills these gaps by detecting spelling patterns in low-fluency readers.

In [Nicolao et al. 2018], the authors proposed a system for assessing children's reading skills by detecting fluency and pronunciation errors. They used a lightly supervised approach to acoustic modeling based on WFST to model specific errors such as repetitions, substitutions, and deletions observed in recordings. [Montoya Gomez et al. 2025] described a model to detect miscues in children's oral reading using an ASR system with phonemic targets, combined with a WFST to model the pronunciation lexicon. Their proposed WFST construction handles multiple pronunciations for a given word. This system was evaluated on a corpus of Dutch-speaking primary school children and outperformed previous results on the same evaluation set. [Yılmaz et al. 2014] extended a two-layered speech recognition architecture (FLaVoR) for automatic reading assessment by incorporating a phone confusion model. This model allows for flexible decoding by considering typical phone substitutions, deletions, and insertions, aiming to improve reading miscue detection, and showed improved performance on the CHOREC database. However, most of the research has been on detecting miscues in a coarse manner, and do not attempt to investigate how well a child has read within that low-fluency class. In this work, we close this gap, first focusing on spelling assessment. In the Brazilian Portuguese setting, to our knowledge, our work is the first to take the northeastern spelling into account.

The authors of [Jain et al. 2023] explored various pretraining and finetuning configurations of Wav2Vec2 for self-supervised learning to improve child ASR, finding that finetuning with even small amounts of child speech data significantly boosts performance on child speech compared to models finetuned only on large adult datasets. [Block Medin et al. 2024] compared Wav2Vec2, HuBERT, and WavLM for phoneme recognition in French children's speech, noting that HuBERT and WavLM performed better than Wav2Vec2. [Gao et al. 2024] investigated pretrained models, including Wav2Vec2, for Dutch child speech recognition and reading miscue detection. They found Wav2Vec2 showed the highest recall for miscue detection, although another model, Whisper, had better precision. Similarly, in the context of Brazilian Portuguese, [Ferreira et al. 2022] compared a supervised TDNN model trained on child speech with a self-supervised Wav2Vec2 model trained on adult speech for assessing children's reading fluency. They found that while the standard Wav2Vec2 model performed poorly, a version augmented with a task-specific language model (Wav2Vec2-lm) achieved a Word Error Rate approximately half that of the TDNN model. Their results showed that for the constrained task of reading a known text, a powerful language model could help a model trained on adult speech outperform a specialized model trained on child speech. In this work, we show that a XSLR model finetuned on child speech can be used in another context, that of spelling detection and assessment.

4. Materials and Method

This section details the three core components of our methodology. First, we fine-tuned a pre-trained XLSR Wav2Vec2 model on 26 hours of children reading pseudo-words.

Second, we developed a spelling dataset from 2000 annotated recordings of second-graders, capturing four distinct reading modalities and accounting for pronunciation variants. Third, we constructed and compared three Weighted Finite-State Transducers (WFSTs) designed to classify these reading modalities by mapping them to unique paths, incorporating features like self-correction loops and an "unk" token for unknown sounds.

4.1. Wav2Vec2 finetuning

We fine-tuned a pre-trained XLSR model using a manually annotated corpus of children reading a list of *pseudo-words* (nonexistent words that nonetheless conform to Brazilian Portuguese phonotactics). Our fine-tuning dataset comprises roughly 26 hours of audio, partitioned into 20.8 hours for training and two 2.6 hour subsets for validation and evaluation.

Training proceeded for 100 epochs. On the evaluation portion of the pseudo-word dataset, the model achieved a phone error rate (PER) of 0.10. To assess generalization, we then tested the fine-tuned model on a separate 14 hour corpus of children reading actual Portuguese words (completely disjoint from the 26 hour pseudo-word set). In this out-of-domain test, the model yielded a PER of 0.24.

4.2. Dataset construction

Correct	Spelled	Spelled + Syllabic	Syllabic
gota	[ʒɛ] [ɔ] [te] [a]	[ʒɛ] [ɔ] [gɔ] [te] [a] [ta]	[gɔ] [ta]
pilotu	[pe] [i] [ɛlɪ] [ɔ] [te] [ɔ]	[pe] [i] [pi] [ɛlɪ] [ɔ] [lɔ] [te] [ɔ] [tɔ]	[pi] [lɔ] [tɔ]

Table 1. Expected IPA transcriptions of the words *Gota*, and *Piloto* under different reading modalities: correct, spelled, spelled with syllabification, and syllabic.

Our spelling dataset consists of audio recordings of children reading a list of 60 words in order. Audio recordings were collected from a large-scale Brazilian fluency assessment conducted in 18 Brazilian states, with the aim of assessing second-year public elementary school students. Two linguists manually labeled the reading modality for all 2000 recordings, and those reading modalities were used to construct paths in our FSTs, which include paths for different reading modalities: correct reading, syllabic reading, spelled plus syllabic reading, and spelled reading. Additionally, for each of the 60 words, we transcribed the expected pronunciation for each modality. For instance, for the words *gota* (droplet), and *piloto* (pilot) the annotated transcriptions can be found in Table 1.

While we included only canonical forms for each modality, children's actual productions often deviated from these templates. In particular, they frequently produced hybrid forms that mixed different reading strategies. For example, instead of the full spelled plus syllabic reading [ʒɛ] [ɔ] [gɔ] [te] [a] [ta], a child might produce [gɔ] [te] [a] [ta]. These mixed forms were not explicitly modeled in this work.

4.3. FSTs construction

In our FSTs, each reading modality corresponds to a distinct path through the transducer, with sound parts representing state transitions. The collection of reading modalities of each word give rise to a "kernel" in the FST (see Figure 1). There are no interactions

Phone	Alternative phone	Northeastern variation
ɛ	e	
ɔ	o	
ẽmɪ	emɪ	me
ẽnɪ	enɪ	ne
ʒe	ge	
ɛlɪ		le
ʒɔta		ʒi
ɛxɪ		xe
ɛsɪ		si
ɛfɪ		fe

Table 2. Phone mapping employed in this work. For each word in out dataset in which one of the phones in the Phone column occur, we create a variant of that word, with the phone replaced by the phones in that row. Enclosing brackets were omitted.

between the word kernels, so the overall transducer is linear and sparse. During word reading classification, we prioritize the highest level decoding class.

Our FSTs include several key features: (i) backward transitions to accommodate self-corrections, which are common in our child reading data; (ii) an ε-transition for skipping an word; (iii) support for multiple valid pronunciations of letters, words and syllables; (iv) an emission matrix modified to explicitly model unknown sounds, as described in the following.

Since our acoustic model was not originally trained to recognize unknown sounds, we implemented the following heuristic to provide this capability. Let P be the emission matrix of our acoustic model, where each column $P(\cdot, t)$ represents a probability distribution over vocabulary tokens at time t. We define:

$$p_{\max}(t) = \max_i P(i, t).$$

To detect unknown sounds, we modify the emission matrix by inserting probability mass into an "unk" token. When $P(\varepsilon, t) < 0.8$ (indicating that the model recognized some sound), we set the probability of the unknown token as:

$$P(\text{unk}, t) = \begin{cases} P(\varepsilon, t) & \text{if } p_{\max}(t) > P(\varepsilon, t), \\ 0.8 \cdot p_{\max}(t) & \text{otherwise.} \end{cases}$$

We then set $P(\varepsilon, t) = 0.001$. These values 0.8, 0.8, and 0.001 were determined through hyperparameter tuning conducted on a validation subset of the data. We performed a grid search over a range of candidate values for each parameter and selected the configuration that maximized F1-score for the spelling class.

Although our main goal is to detect spelled out readings, we still need to encode the other modalities paths in each WFST. Limiting the decoder to spelling arcs would force every input through the spelling route, which could inflate recall at the cost of

precision by mislabeling fluent speech as spelled segments. Including alternative reading routes allows the transducer to match smooth or partially correct utterances to their intended trajectories, reserving the spelling channel only for true spelled out patterns. This balanced design upholds precision by keeping non-spelling vocalizations from being wrongly classified as spellings.

4.4. Classification

To classify the reading modality $k \in V$ for each word i, we first define E_{ik} as the set of allowed pronunciations of word i for reading modality k, where $|E_{ik}| \geq 1$ accounts for multiple possible correct pronunciations, and V is an index set over the reading modalities. Let x_i be the output sequence from the FST for word i. We define an indicator function:

$$Q_k(x_i) = \begin{cases} 1, & \exists\, e \in E_{ik} \text{ such that } e \text{ is a substring of } x_i, \quad k \neq \eta, \\ 1, & \exists\, e \in E_{i\eta} \text{ such that } e = x_i, \quad k = \eta, \\ 0, & \text{otherwise}, \end{cases}$$

where η denotes the spelling modality, which requires an exact match rather than a substring match. The final classification is determined by scanning the modalities in decreasing order of reading fluency and selecting the first modality that satisfies our criteria:

$$\hat{k}_i = \min_{k \in V} \{k : Q_k(x_i) = 1\}.$$

This approach prioritizes more fluent reading modalities when multiple classifications are possible. If $Q_k(x_i) = 0$ for all reading modalities k, we classify that word reading as "Wrong".

The output from our WFSTs looks like: [gɔ_0, ta_0, kaneta_1, pilotu_2, dadu_3, ka_4, bi_5, dɛ_6, (...)]. We parse the output by the number suffix, and each x_i is sent to the classifier. For the provided example, $x_0 =$ [gɔ, ta], $x_2 =$ [pilotu]. We then scan the allowed pronunciations E_{ik} of words $i = 0$ (*gota*) and $i = 2$ (*piloto*) (see Table 1), and classify them as Syllabic and Correct, respectively.

4.5. FST Variants for Experimental Comparison

To evaluate which modeling choices contribute most to spelling classification, we constructed three increasingly sophisticated transducers: G0, G1, and G2.

G0 is a minimal, linear and sparse FST that includes only the canonical reading paths for each modality (see Figure 1). Specifically, each word kernel contains a single IPA sequence per modality, with no alternate pronunciations. There is no explicit "unk" token. Backward self-corrections are allowed only at the full-word level (no intra-word loops), and there are no word-skip ε transitions.

G1 builds on G0 by adding support for multiple pronunciations of letters via the mappings described in Section 4.1. In this variant (see Figure 2), each kernel includes all valid alternative phone arcs, which helps tolerate child-specific phonetic variations. We also introduce ε-transitions for skipping entire words. G1 still does not model an explicit unknown sound ("unk") token in the emission matrix.

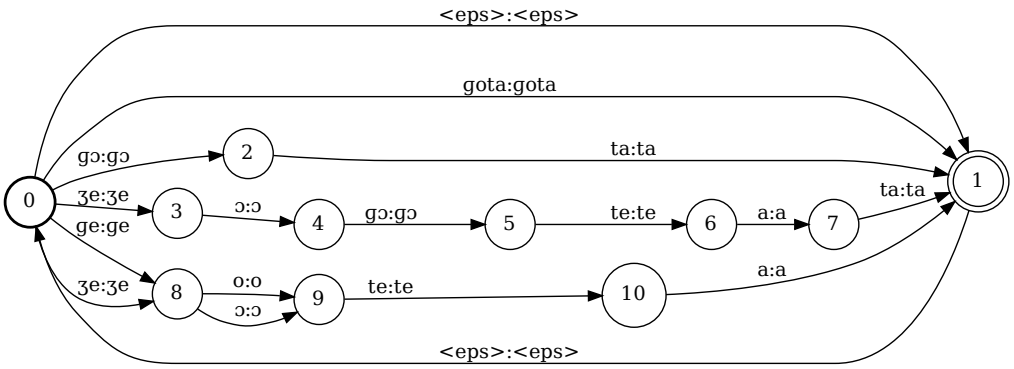

Figure 2. Word *gota* kernel in grammar G1.

G2 retains all features of G1 and further extends the transducer to handle unknown sounds and intra-word corrections more robustly (see Figure 3). First, we insert probability mass into an explicit "unk" token within the emission matrix (Section 4.3). Second, self-correction loops are introduced to both spelled and spelled plus syllabic arcs, allowing the FST to accommodate mid-word corrections during spelling. These enhancements address the common phenomenon of children producing completely unexpected sounds (e.g., hesitations or sniffs) and self-correcting within words.

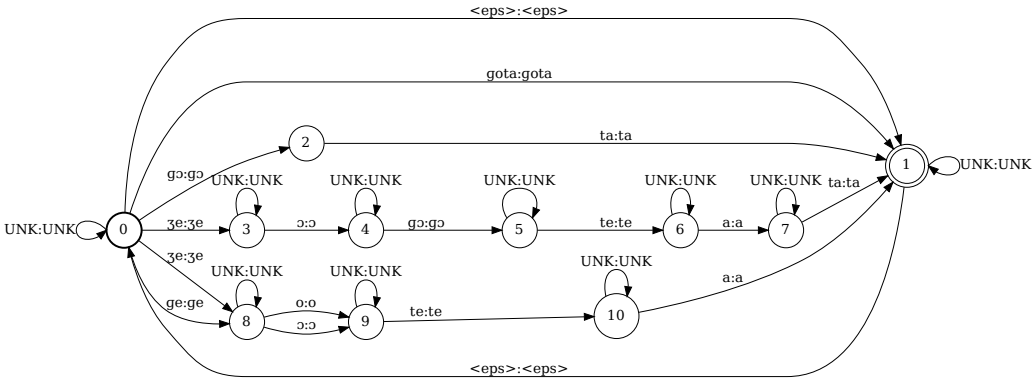

Figure 3. Word *gota* kernel in grammar G2.

5. Experiments and Results

We evaluated G0, G1, and G2 on our test set of 2000 child readings on a set consisting of 2,000 child speech recordings. The evaluation results, summarized in Table 3, reveal a clear trade-off: as the grammars become more complex from G0 to G2, recall consistently increases at the expense of precision. The introduction of phone variants in G1 yields a significant recall improvement (from 0.7786 to 0.8701) while only moderately decreasing precision, resulting in the highest F1-score of 0.7129. The further addition of unk-loops

Grammar	Accuracy	Precision	Recall	F1-Score
G0	0.5971	0.6523	0.7786	0.7099
G1	0.6358	0.6037	0.8701	0.7129
G2	0.6320	0.5598	0.9060	0.6920

Table 3. Overall performance metrics (accuracy, precision, recall, and F1-score) for each grammar variant on the spelling classification task.

in G2 pushes recall to its peak at 0.9060 by capturing more varied utterances, but this flexibility causes a steeper drop in precision to 0.5598. While overall accuracy remains comparable between G1 and G2, the more balanced G1 grammar achieves the best F1-score, highlighting the precision-recall trade-off inherent in our more complex models.

To quantify these shifts, we present in Table 4 the raw counts of true positives, false positives, and false negatives for each grammar's Spelling detections. The reduction in false negatives from 271 in G0 to 115 in G2 directly drives the observed recall improvement (0.7786 to 0.9060), meaning that G2 detects 156 more actual spelled readings than the baseline. Conversely, the number of false positives increases from 508 for G0 to 872 for G2, accounting for the decline in precision (0.6523 to 0.5598). In other words, G2 incorrectly labels 364 additional non-spelling excerpts as spelled compared to G0, which highlights a bias toward over-classification of the Spelling class.

Grammar	True Positives	False Positives	False Negatives
G0	953	508	271
G1	1065	699	159
G2	1109	872	115

Table 4. Confusion matrix breakdown for the Spelling class, showing how many spelled readings were correctly identified, how many non-spelling readings were mislabeled as spelling, and how many spelled readings were missed.

The primary factor underlying these performance shifts appears to be the inclusion of the unk-loops in G2's spelled-reading path (depicted in Figure 3). These loops are intended to absorb acoustic uncertainty but they also inadvertently admit a broader range of non-spelling phonatory events (e.g., disfluent speech or background noise). As a result, non-spelling utterances with elongated phonation or atypical prosody can satisfy the relaxed constraints of G2's grammar, yielding more false positives. In contrast, G1, which only incorporates phone variants, strikes a more balanced trade-off: it substantially reduces false negatives (from 271 to 159) without inflating false positives to the same extent as G2 does.

In summary, introducing phone variants (G1) yields a substantial recall improvement over the baseline (G0) with only a moderate precision penalty, while the further addition of unk-loops arcs in G2 pushes recall even higher at the cost of accepting more non-spelling vocalizations as spelled. Given the relative importance of recall in our intended diagnostic context, G2's configuration is deemed preferable.

6. Conclusion

This study demonstrated that Weighted Finite-State Transducers provide an effective framework for detecting oral spelling strategies in young readers. Our best-performing model achieved high recall by accommodating regional pronunciation variants and decoding uncertainties, offering valuable diagnostic insights beyond traditional assessment metrics. In the following subsections, we discuss the pedagogical implications of these results, address the limitations of our current approach, and outline promising directions for future work.

6.1. Discussion

Our results confirm that a WFST-based approach can successfully model and detect specific oral spelling strategies employed by young readers of Brazilian Portuguese. By integrating phoneme variants (including regional "northeastern alphabet" letter-naming conventions), and explicit unk-loops to absorb unexpected sounds, our most advanced grammar (G2) achieves a recall of 0.9060 for the Spelling class. Given the context of educational screening (identifying students in risk of poor learning outcomes), where it is generally considered more important to prioritize sensitivity (recall) over specificity (precision) in order to reduce the risk of overlooking children who might need support [Glover and Albers 2007], we opted to favor recall (recognizing that this choice could come at the expense of some precision). Pedagogically, our approach provides actionable diagnostic information beyond the traditional binary "correct/incorrect" word-level scoring. Such detailed diagnostic distinctions enable educators to tailor interventions to each student's specific decoding profile.

Beyond the specific metrics, our work emphasizes that automatic oral reading assessments must move beyond coarse-grained "fluent vs. non-fluent" labels. By capturing the continuum of decoding strategies WFST-based systems can better reflect the pedagogical reality of early readers in Brazil.

6.2. Limitations and future work

Despite the promising results, several limitations of our study warrant discussion and motivate avenues for future research.

Our emphasis on recall inevitably permits a higher rate of false positives, where wrong or partially correct spelled readings may be misclassified as spell-outs. While acceptable in a screening context, this trade-off could lead to unnecessary follow-up assessments if used without subsequent precision-focused validation.

We did not perform fine-grained phoneme-level error annotation on the actual recordings, making it difficult to determine precisely which phoneme variants or unknown sounds caused misclassifications. Although our phone-variant mapping mitigates some variability, there remain cases where a child's nonstandard phonetic realization falls outside our predefined mapping or where the acoustic model's emission matrix fails to assign sufficient probability mass to the correct phoneme. A detailed phoneme-level labeling effort would help to systematically identify the most problematic phones for spelling detection and refine both the mapping table and the FST transitions.

While our primary focus was on spelling detection, children produce a variety of miscues during oral reading. Existing literature [Ávila et al. 2009] [Luna et al. 2025]

identifies phenomena such as "prolonged syllabification," "phoneme blending errors," and "word-skipping hesitations," each of which may require distinct modeling strategies within the WFST framework. For example, a child might begin to spell "piloto" as "[pɐ] [i]" then switch to a correct phonemic rendering for the remainder of the word. Modeling such hybrid miscues in a fine-grained way would improve diagnostic specificity but also complicate the transducer design.

Although our classification metrics indicate that G2 yields high recall for spelling detection, the ultimate goal is to inform pedagogical practice. In practice, WFST-based decoding feedback could be embedded into existing large-scale fluency assessments or classroom diagnostic tools, automatically generating reports that highlight whether a child predominantly spells, syllabifies, or blends phonemes. Such information can orient teachers toward targeted strategies (for example, focusing on phoneme blending when spelling predominates, or reinforcing letter–sound correspondences when syllabification errors are frequent). We have not yet measured whether and how providing this feedback leads to improved instructional decisions or better student outcomes. Future experiments should integrate our system into classroom workflows, collect teacher feedback on the usability of these reports, and evaluate whether interventions informed by fine-grained decoding profiles accelerate reading acquisition compared to standard fluency assessments.

Finally, it is important to note that our experiments were conducted using a single WFST architecture, selected after preliminary tests indicated superior performance relative to other simple configurations. We did not exhaustively explore alternate transducer topologies and it remains an open question whether these could yield further gains in precision or recall. Future research should systematically investigate a broader range of FST architectures to determine which designs most effectively balance robustness to acoustic variability with the pedagogical need for high recall in early decoding detection.

References

Ávila, C. R. B. d., Kida, A. d. S. B., Carvalho, C. A. F. d., and Paolucci, J. F. (2009). Tipologia de erros de leitura de escolares brasileiros considerados bons leitores. *Pro. Fono.*, 21(4):320–325.

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

Block Medin, L., Pellegrini, T., and Gelin, L. (2024). Self-supervised models for phoneme recognition: Applications in children's speech for reading learning. In *Interspeech 2024*, pages 5168–5172, ISCA. ISCA.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition.

Ferreira, A. L., Silva, C., de Assis, E., and de Souza, J. (2022). Avaliação de modelos para reconhecimento automático de fala aplicados para identificação da qualidade de leituras em voz alta de narrativas breves. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 895–907, Porto Alegre, RS, Brasil. SBC.

Gao, L., Tejedor-Garcia, C., Strik, H., and Cucchiarini, C. (2024). Reading miscue detection in primary school through automatic speech recognition. In *Interspeech 2024*, pages 5153–5157, ISCA. ISCA.

Glover, T. A. and Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, 45(2):117–135. Universal Screening for Enhanced Educational and Mental Health Outcomes.

Gough, P. B. and Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1):6–10.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Guo, C., Lian, J., Zhou, X., Zhang, J., Li, S., Ye, Z., Park, H. J., Das, A., Ezzes, Z., Vonk, J., Morin, B., Bogley, R., Wauters, L., Miller, Z., Gorno-Tempini, M., and Anumanchipalli, G. (2025). Dysfluent wfst: A framework for zero-shot speech dysfluency transcription and detection.

Hulme, C. and Snowling, M. J. (2013). Learning to read: What we know and what we need to understand better. *Child Development Perspectives*, 7(1):1–5.

Jain, R., Barcovschi, A., Yiwere, M. Y., Bigioi, D., Corcoran, P., and Cucu, H. (2023). A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. *IEEE Access*, 11:46938–46948.

Kouzelis, T., Paraskevopoulos, G., Katsamanis, A., and Katsouros, V. (2023). Weakly-supervised forced alignment of disfluent speech using phoneme-level modeling.

Luiz Gonzaga and Zé Dantas (1987). Abc do sertão.

Luna, A. S., Machado-Lima, A., and Nunes, F. L. S. (2025). Identification and classification of speech disfluencies: A systematic review on methods, databases, tools, evaluation and challenges. *Journal of the Brazilian Computer Society*, 31(1):154–173.

Mohri, M., Pereira, F., and Riley, M. (2008). *Speech Recognition with Weighted Finite-State Transducers*, pages 559–584. Springer Berlin Heidelberg, Berlin, Heidelberg.

Montoya Gomez, G. M., Ghesquiere, P., and Van hamme, H. (2025). Reading proficiency assessment using finite-state transducers. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers*, ICETC '24, page 332–338, New York, NY, USA. Association for Computing Machinery.

Neubig, G., Akita, Y., Mori, S., and Kawahara, T. (2012). A monotonic statistical machine translation approach to speaking style transformation. *Computer Speech Language*, 26(5):349–370.

Nicolao, M., Sanders, M., and Hain, T. (2018). Improved acoustic modelling for automatic literacy assessment of children. In *Interspeech 2018*, pages 1666–1670.

Rocha, C., Mello, R., and Souza, J. (2024). Avaliação de fluência leitora em língua portuguesa: primeira experiência com uso em larga escala de inteligência artificial. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 3075–3084, Porto Alegre, RS, Brasil. SBC.

Silva, C., Ferreira, A. L., de Assis, E., and de Souza, J. (2022). Definição de heurística para identificação automática da fluência em leitura de crianças em fase de

alfabetização. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 39–50, Porto Alegre, RS, Brasil. SBC.

Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21:360–407.

Yılmaz, E., Pelemans, J., and Van hamme, H. (2014). Automatic assessment of children's reading with the flavor decoding using a phone confusion model.