Automatic Speech Recognition for Child Reading: A Phonemic Approach using Isolated Words in Brazilian Portuguese

Aline N. Rodrigues and Carlos H. C. Ribeiro
Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos, SP, Brazil
{alineanr,carlos}@ita.br

Abstract

Automatic assessment of reading in children who are learning to read is challenging due to the lack of data and the high variability of children's speech. This work investigates the improvement of Automatic Speech Recognition (ASR) models for the analysis of reading decoding of isolated words in Brazilian Portuguese. We propose a methodology based on fine-tuning Wav2Vec2.0 models, with a paradigm transformation from orthographic to phonemic transcription. Using a novel corpus of 5,400 audio word samples from children in the 2nd and 3rd grades of Elementary School, we compare pre-trained models in Portuguese and multilingual. Results reveal that the phonemic approach, combined with fine-tuning strategies, data augmentation, and adapted tokenization, significantly reduces the Phoneme Error Rate (PER). This overcomes the limitations of commercial tools and validates the use of ASR for the detailed diagnosis of decoding errors and phonological acquisition.

1 Introduction

The ability to read is a fundamental skill for cognitive and social development (Navas et al., 2009; Andrade et al., 2019). In the early years of schooling, decoding ability, i.e., the ability to transform graphemes into phonemes, is a key predictor of future reading fluency. Educators and speech therapists conduct a systematic assessment of this skill through individual read-aloud lessons. During these assessments, the specialist must listen carefully to the child, manually control the reading time, and record specific decoding errors, such as substitutions, omissions, or hesitations of phonemes, to calculate metrics such as reading accuracy or words correct per minute (WCPM). However, as this process is highly individualized, open to subjective interpretations, and takes a long time, it is difficult and not very scalable for regular monitoring in large education systems.

Automatic Speech Recognition (ASR) is a potentially promising solution. However, applying ASR to children's speech, particularly for Brazilian Portuguese (BP), has faced serious challenges. First, there is a **lack of data**: public datasets of children's speech are scarce (Sriram et al., 2022; Zhu et al., 2022; Molenaar et al., 2023; Duan, 2023), which limits supervised training. Second, the inadequacy of adult models: models trained with adult speech fail to handle the acoustic variability (higher fundamental frequency, distinct formants) and pronunciation errors typical of children (Yılmaz et al., 2014; Johnson et al., 2023; Rodrigues et al., 2023). In addition, commercial tools (e.g., Google, Amazon) often operate as "black boxes," providing orthographic transcriptions that mask subtle decoding errors, missing essential information for pedagogical diagnosis.

This article focuses on the challenge of reading isolated words. In contrast to reading narrative texts, where the semantic context helps predict the next word, reading isolated words assesses the child's competence in decoding without the support of contextual elements.

It is our hypothesis that ASR models based on purely phonemic transcription, adapted using fine-tuning on specific data, are more accurate than orthographic models in representing the real situation of children's speech production.

The contributions of this work include:

1. The corpus curation and phonemic annotation of isolated words spoken by children learning to read;

2. A customized tokenization methodology to handle complex phonemes from the SAMPA standard in Wav2Vec2.0;

3. A three-phase training strategy with acoustic data augmentation;

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 974–979

4. A detailed analysis to demonstrate the advantage of BP pre-trained models compared to massive multilingual models for this task.

2 Related Work

Literature research indicates that end-to-end ASR models, such as Wav2Vec2.0 (Baevski et al., 2020), provide the state-of-the-art performance by learning acoustic representations from raw audio. Adaptation to child speech (Child ASR) in low-resource languages, however, represents a challenge.

Children's speech displays higher variability and lower articulatory accuracy than adult speech. Studies such as (Rodrigues et al., 2023) and (Johnson et al., 2023) explore the use of Wav2Vec2.0 for children speech, but often focus on continuous speech or use adult data to mitigate scarcity. (Yılmaz et al., 2014) proposed phonetic confusion models to handle mispronunciations, but recent approaches to direct fine-tuning on phonemes for BP have been little explored.

In contrast to approaches that aim only to optimize the word error rate (WER), this work prioritizes the phoneme error rate (PER)[1]. In an educational context, knowing *how* the child made the mistake (e.g., changing /v/ to /f/) is more valuable than just knowing that the word is wrong.

3 Methodology

3.1 Children's Speech Corpus

A specific corpus was created using recordings of 208 children in the 2nd and 3rd grades of elementary school. Data collection took place in the cities of Belo Horizonte (MG) and São José dos Campos (SP), covering the Southeastern Brazilian dialect. The collection followed rigorous ethical protocols. The reading source consisted of a list of 24 high-frequency words developed for reading assessment in Brazilian Portuguese (Lúcio et al., 2018) (e.g., *farta*, *nublado*, *treze*, *enxuto*, and *famoso*).

Preprocessing and Alignment: For the precise extraction of each word, a forced alignment approach was employed using the Whisper model (Radford et al., 2023) (large-v2), through the *stable-ts* library (Jian, 2025). To ensure quality, a human verification process was conducted on a data subset. Initially, the algorithm produced imprecise cuts, resulting in phoneme loss. We iteratively adjusted the padding parameters and repeated the validation until the verified samples showed no clipping, ensuring the integrity of the word boundaries. The final corpus comprises 5,400 samples (approximately 2.1 hours).

Stratification: The dataset was split into training (60%), validation (20%), and testing (20%). To ensure representation and prevent bias, double stratification was applied based on:

1. **Grade:** Ensuring an equal proportion of 2nd and 3rd grade students.

2. **Score:** Classification based on the child's percentage of correct words, splitting the sample into performance quartiles.

3.2 Phonemic Approach and Tokenization

To address the decoding analysis, we decided to train the ASR model to predict phonemes instead of graphemes. The orthographic transcriptions (ground truth) were converted to phonemic representations using the FalaBrasil project transcriptor (Siravenha et al., 2008). Although IPA (International Phonetic Alphabet) is the standard for multilingual ASR, we opted for the SAMPA standard to maintain direct compatibility with FalaBrasil, which is currently the most robust G2P tool specialized for Brazilian Portuguese. Converting its native output to IPA could introduce noise or mapping errors, so we prioritized the reliability of the phoneme generation over universal notation.

The Vocabulary Challenge: There was a technical challenge in getting this approach to work with Wav2Vec2.0. The default tokenizer in the transformers library is based on single characters, but the SAMPA standard represents some Portuguese phonemes with multiple characters. For example, the phoneme /$\tilde{a}$/ is represented as 'a$\tilde{}$' and the phoneme /$\int$/ (x) as 'tS'. Using the standard tokenizer breaks 'tS' into the tokens 't' and 'S', resulting in an incorrect phonemic representation.

To overcome this issue, we implemented a customized mapping (Table 1) where each multi-character phoneme was converted to a unique Unicode token, ensuring one-to-one alignment between the output *token* and the acoustic target.

[1]Both WER and PER are calculated based on the Levenshtein distance $(S + D + I)/N$, where S, D, and I represent the number of substitutions, deletions, and insertions, respectively, and N is the total number of words (for WER) or phonemes (for PER) in the reference transcription.

Phoneme	SAMPA	New Token
/t∫/ (bye)	tS	Č
/d3/ (day)	dZ	Š
/$\tilde{a}$/ (wool)	a∼	Ã
/$\tilde{e}$/ (well)	e∼	$\tilde{E}$
/$\tilde{w}$/ (hand)	w∼	$\tilde{W}$

Table 1: Example of mapping multi-character SAMPA phonemes to single tokens.

3.3 Experiment Setup

We explored two groups of pre-trained base models to measure the impact of this research domain:

1. **Expert in BP (Adult):** CORAA (Casanova, 2022) and Grosman-XLS-R-1B (Grosman, 2022). The assumption is that prior exposure to Portuguese phonetics would improve adaptation, even with adult speech.

2. **Multilingual:** XLSR-53 and XLS-R-300M/1B (Babu et al., 2021). The hypothesis is that vast linguistic diversity would bring robustness to the variability of children's speech.

To provide context, it is important to note that all of these selected base models were originally trained for orthographic transcription (character to word prediction). To adapt them to our task, we discarded their original tokenizers and linear classification headers. During fine-tuning, we replaced them with our custom phonemic tokenizer. The purpose was to benefit from the robust acoustic-linguistic representational space already learned by these models, including the phonetic characteristics of Portuguese in BP and to redirect the output to predict phonemes as opposed to graphemes.

Training Configuration: Hyperparameters were standardized across experiments based on preliminary tests. We used the CTC loss function and AdamW optimizer with a learning rate of $3e^{-5}$ (linear schedule, no warmup). Dropout was set to 0.1 (attention hidden) and 0.07 (encoder), with a weight decay of 0.01. To prevent overfitting, early stopping was applied with a patience of 20 epochs, monitoring validation loss.

Phased Training Strategy: To evaluate the impact of data augmentation (ablation study), we compared models with and without the augmentation pipeline. For the augmented models, we used a three-phase protocol:

- **Phase 1 (Initial Adaptation):** 30 epochs using only real data. The objective of this phase is to allow the pre-trained base model to initially adapt to the primary acoustic and linguistic characteristics of the children's speech corpus.

- **Phase 2 (Generalization):** 30 epochs training from the Phase 1 checkpoint, using a combined dataset of real and augmented data. The augmentations included Time Stretch ($\pm 5\%$), Pitch Shift (± 1 semitone), and Additive Noise. To preserve speech intelligibility, a random combination of no more than two of these techniques was applied simultaneously to each original audio sample. This phase exposes the model to greater acoustic variability to improve its robustness.

- **Phase 3 (Refinement):** 30 epochs training from the Phase 2 checkpoint, only real data. Unlike Phase 1, the goal of this final refinement step is to recalibrate the already robust model back to the natural speech distribution, mitigating any potential artificial bias introduced by the synthetic augmented data from the previous phase.

We also reset the weights of the last three Transformer layers to facilitate domain adaptation (Pasad et al., 2021). Note that the custom tokenization (Table 1) is a structural requirement for the phonemic approach and cannot be ablated.

4 Results and Discussion

4.1 Baseline vs. Fine-Tuning

We conducted a baseline evaluation using pre-trained models on the standard orthographic task (character/word prediction). The performance was evaluated using WER. For the proposed phonemic approach, we evaluated PER.

Results are shown in Table 2. The baseline models, even those with 1 billion parameters (Grosman-1b), performed unsuccessfully for diagnostic purposes, with a WER greater than 0.50 and an estimated PER of 0.1771. This suggests that models trained on adults face challenges in segmenting and recognizing isolated children's speech without context.

The use of phonemic fine-tuning produced a significant improvement. The best fine-tuned model (CORAA-aug) achieved a PER of 0.0437, a relative error reduction of 75.3% when comparing this

result directly to the best baseline PER achieved before fine-tuning (Grosman-1b Adult, PER of 0.1771). The statistical significance of this improvement was assessed using a Wilcoxon test on the paired data. Specifically, we calculated the PER for each audio sample in the test set for both the baseline and fine-tuned models. The paired test confirmed that the performance gain is statistically significant ($p < 0.05$)

Model (Condition)	Metric	Value
Baseline (Orthographic)		
CORAA (Adult)	WER	0.6045
CORAA (Adult)	PER	0.2187
Grosman-1b (Adult)	WER	0.5410
Grosman-1b (Adult)	PER	0.1771
Fine-tuning (Phonemic)		
CORAA (BP)	PER	0.0473
CORAA-aug (BP)	**PER**	**0.0437**
Grosman-1b (BP)	PER	0.0605
Grosman-1b-aug (BP)	PER	0.0518
Facebook-1b-aug (Multi)	PER	0.0590

Table 2: Performance comparative: Orthographic Baseline vs. Phonemic Fine-tuning. (aug = with data augmentation).

For the ablation study, comparing the fine-tuning method (CORAA) with the augmented one (CORAA-aug) shows a reduction in PER from 0.0473 to 0.0437. Although the absolute difference is small, it represents a 7.6% relative error reduction. More importantly, the three-stage training strategy with data augmentation proved to be a valuable regularizer for mitigating overfitting in our low-resource dataset, exposing the model to acoustic variations (e.g., pitch changes and noise) that increase its overall robustness.

4.2 Base Model Effect: BP vs. Multilingual

An important finding of this research comes from comparing specialized Portuguese models with multilingual models. For the task of phonemic recognition of isolated words, the Portuguese-based models (CORAA and Grosman) consistently outperformed their multilingual equivalent (Facebook XLS-R) of the same type.

The CORAA-aug model (300M parameters) was significantly better than Facebook-XLS-R-300M ($p < 0.001$). Even the large multilingual model with 1 billion parameters (PER 0.0590) could not outperform the smaller specialized model (CORAA, PER 0.0437). This means that for accurate phonetic decoding in a low-data environment, prior alignment with the phonotactics and spectral characteristics of the target language (acquired during adult pre-training) is likely more helpful than just having a high generalization ability in a multilingual setting.

4.3 Performance by School Grade

Although the speech of 2nd-grade children (approx. 7-8 years old) is typically more challenging for ASR systems than that of older students due to a higher incidence of phonological simplification processes and greater variability in segment duration, our fine-tuning approach successfully overcame these difficulties. When analyzing the performance of the best model (CORAA-aug) by grade level, the PER for the 2nd grade was 0.041, compared to 0.046 for the 3rd grade. This demonstrates that phonemic fine-tuning was highly effective in capturing the specific phonetic characteristics of younger readers, achieving a lower error rate for this group compared to the 3rd graders.

This is a promising result, as it shows that the fine-tuning technique was able to adapt the model to the more challenging acoustic characteristics of second graders, enabling the tool to be used for the age group where early diagnosis is most critical.

4.4 Qualitative Analysis of Reading Deviations

The phonemic approach changed the interpretation of incompatible transcriptions in a significant way. While the baseline produced ASR "hallucinations" (prediction of words with no phonetic relation to the audio), the results of the adjusted model proved to be phonetically close to the actual speech productions of the children.

Expert analysis by speech therapists confirmed that what initially appeared to be "ASR errors" in relation to the target text were, in fact, accurate transcriptions of the children's own phonological and articulatory deviations. We identified two main patterns that reflect language acquisition processes:

1. **Place of Articulation Substitutions:** When a child mispronounces a phoneme, the model captured the exact substitution. Example: Target text "/z e/" (zê) $\rightarrow$ Actual child utterance and ASR Prediction "/Z e/" (gê). The model accurately detected the change from an alveolar to a post-alveolar fricative.

2. **Manner of Articulation:** Example: Target text "/p i/" → ASR prediction "/s i/". Experts confirmed that children typically produce plosives with very weak release energy. The model's prediction reflects this precise acoustic reality (a softened plosive, similar to a fricative) in contrast to a random fail.

These results show that the adjusted ASR not only "fails" when the reading is incorrect but also acts as a diagnostic mirror. Experts have confirmed that capturing these small phonetic deviations is significantly more useful for planning pedagogical interventions than the binary "wrong word" flag provided by standard orthographic ASRs.

4.5 Interpretability of Latent Space

To verify if the model learned phonetic distinctions, we analyzed embeddings from the Transformer's final layer, as last layers best represent task-specific spaces after domain adaptation (Pasad et al., 2021). Applying UMAP and Spherical K-Means (Grosz et al., 2023), we observed the latent space evolution (Figure 1).

To formally evaluate class purity, we used the Homogeneity metric (Rosenberg and Hirschberg, 2007). Initially, the pre-trained model showed diffuse clusters with a low Homogeneity score of 0.35. Post-fine-tuning, this score surged to 0.94. Furthermore, the resulting clusters are phonetically coherent: the model spontaneously grouped front mid vowels (/e/, /E/, /ẽ/) and voiced sibilants (/z/, /Z/) in distinct regions. This proves the fine-tuning successfully restructured the acoustic knowledge specifically for child vocal physiology.

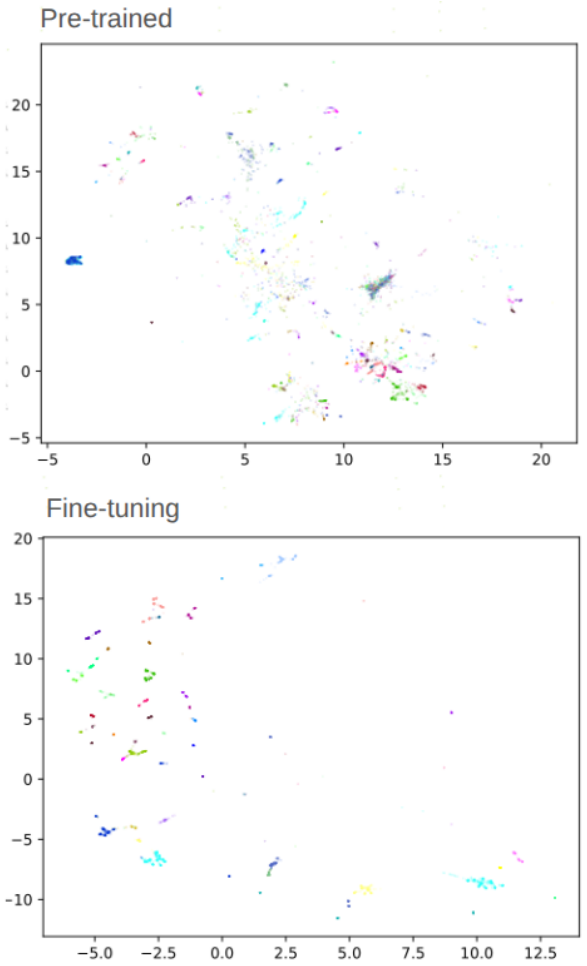

Figure 1: UMAP visualization of phonemic embeddings. (A) Pre-trained: diffuse and mixed clusters. (B) Post-fine-tuning: distinct and dense phonetic clusters.

5 Conclusions

This research shows that automatic transcription of isolated words for children in the literacy stage requires a different methodological approach than that used for continuous speech from adults. The use of fine-tuning in Wav2Vec2.0 models, combined with a pure phoneme representation and appropriate tokenization treatment, overcame the limitations imposed by the lack of data.

The main conclusions are:

1. **Phonemic Superiority:** For diagnostic assessment of isolated reading, predicting phonemes is methodologically more adequate than predicting graphemes.

2. **Impact of Language Pre-training:** Models pre-trained in the target language (BP), such as CORAA, converge better and produce lower error rates (PER 0.04) than multilingual models, suggesting that prior specialization is essential in low-resource contexts.

3. **Diagnostic Tool:** The developed model not only transcribes accurately but also generates latent representations consistent with phonology, opening the way for automatic tools for large-scale reading difficulty detection and analysis in Brazilian schools.

For future work, we intend to evaluate the approach using a corpus of pseudowords, allowing us to test the robustness of the model in the absence of semantic bias. In addition, we propose to integrate fluency metrics into the ASR output to automatically identify disfluencies, aiming at building tools that give detailed feedback on the reading process

to educators and speech therapists.

Acknowledgments

The authors would like to thank all those involved in this project from the following institutions: Ministry of Education (MEC); the Technological Institute of Aeronautics (ITA); the Institute of Mathematics and Statistics, University of São Paulo (IME–USP); the Federal University of Minas Gerais (UFMG); and the Santa Casa School of Medical Sciences.

References

Alair Junio Lemes de Andrade, Letícia Correa Celeste, and Luciana Mendonça Alves. 2019. Caracterização da fluência de leitura em escolares do ensino fundamental ii. *Audiology - Communication Research*, 24:e1983.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong an, Naman Singh, Aditya Saraf, Tatiana Likhomanenko, F-aith Sincl, Alexandre Défossez, and 1 others. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint*.

Edresson Casanova. 2022. Wav2vec 2.0 trained with coraa portuguese dataset. `https://huggingface.co/Edresson/wav2vec2-large-xlsr-coraa-portuguese`.

Richeng Duan. 2023. Joint learning feature and model adaptation for unsupervised acoustic modelling of child speech. In *Proc. INTERSPEECH 2023*, pages 5227–5231.

Jonatas Grosman. 2022. Fine-tuned XLS-R 1B model for speech recognition in Portuguese. `https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-portuguese`.

Tamas Grosz, Yaroslav Getman, Ragheb Al-Ghezi, Aku Rouhe, and Mikko Kurimo. 2023. Investigating wav2vec2 context representations and the effects of fine-tuning, a case-study of a finnish model. In *Proc. INTERSPEECH 2023*, pages 196–200.

Jian. 2025. stable-ts: Modifies openai's whisper to produce more reliable timestamps. `https://pypi.org/project/stable-ts/`. Version 2.19.0.

Alexander Johnson, Hariram Veeramani, Natarajan Balaji Shankar, and Abeer Alwan. 2023. An equitable framework for automatically assessing children's oral narrative language abilities. In *Proc. INTERSPEECH 2023*, pages 4608–4612.

Patrícia Silva Lúcio, Hugo Cogo Moreira, Adriana de Souza Batista Kida, Carolina Alvez Ferreira de Carvalho, Ângela Maria Vieira Pinheiro, Jair de Jesus Mari, and Clara Regina Brandão de Avila. 2018. Word decoding task: Item analysis by irt and within-group norms. *Psicologia: Teoria e Pesquisa*, 34:e3437.

Bo Molenaar, Cristian Tejedor-Garcia, Catia Cucchiarini, and Helmer Strik. 2023. Automatic assessment of oral reading accuracy for reading diagnostics. In *Proc. INTERSPEECH 2023*, pages 5232–5236.

Ana Luiza Gomes Pinto Navas, Joana Cecilia Baptista Ramalho Pinto, and Paula Roberta Rocha Dellisa. 2009. Avanços no conhecimento do processamento da fluência em leitura: da palavra ao texto. *Revista da Sociedade Brasileira de Fonoaudiologia*, 14(4):553–559.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Aline Rodrigues, Gabriela Ribeiro, Victor Silva, Wesley Carvalho, Miguel Ramírez, Luciana Alves, Marcelo Finger, Ana Luiza Navas, and Carlos Ribeiro. 2023. AI and reading fluency for brazilian portuguese: A preliminary study. SSRN Preprint.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Ana Siravenha, Nelson Neto, Valquíria Macedo, and Aldebaro Klautau. 2008. Uso de regras fonológicas com determinação de vogal tônica para conversão grafema-fone em Português Brasileiro. *7th International Information and Telecommunication Technologies Symposium*.

Anuroop Sriram, Michael Auli, and Alexei Baevski. 2022. Wav2vec-aug: Improved self-supervised training with limited data. *Preprint*, arXiv:2206.13654.

Emre Yılmaz, Joris Pelemans, and Hugo Van hamme. 2014. Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model. In *Interspeech 2014*, pages 969–972.

Han Zhu, Li Wang, Gaofeng Cheng, Jindong Wang, Pengyuan Zhang, and Yonghong Yan. 2022. Wav2vec-s: Semi-supervised pre-training for low-resource asr. In *Proc. Interspeech 2022*, pages 4870–4874.