



Improving Automatic Speech Recognition for Children's Reading Assessment with Disfluency-aware Language Models

Jazmín Vidal^{1,2}, Luciana Ferrer², Juan Esteban Kamienkowski^{1,2,3}, Pablo Riera²

¹Departamento de Computación, FCEyN, UBA, Argentina

²Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina

³Maestría de Explotación de Datos y Descubrimiento del Conocimiento, UBA, Argentina

jvidal@dc.uba.ar, lferrer@dc.uba.ar, juank@dc.uba.ar, priera@dc.uba.ar

Abstract

We present an approach to improve automatic speech recognition (ASR) quality for children's reading assessment in Spanish by incorporating disfluencies into the decoding process. Existing ASR-based assessment methods use hybrid or end-to-end models, with some methods relying on task-specific lattices or n-gram models for modeling potentially disfluent speech. In this work, we compare both families of approaches, in combination with in-domain fine-tuning of the acoustic model, on a dataset of Spanish-speaking children reading aloud. The task-specific n-gram model is learned on a synthetic dataset of disfluent text generated automatically from the reference text with our *disfluentES* toolkit. Results show that modeling disfluencies improves ASR performance as well as fluency assessment performance.

Index Terms: reading assessment, disfluency detection, speech recognition

1. Introduction

Reading fluency, defined as reading expressively with accuracy and at a good pace, is a strong predictor of reading comprehension and future academic success [1]. Yet, 46% of third-graders in Latin America struggle with it [2]. Manual scoring of reading tests, which typically requires counting correct words, is costly, time-consuming and often subjective. Automatic reading assessment (ARA) could offer a scalable and efficient alternative, enabling fluency measurement and objective error diagnosis.

ARA involves analyzing fluency components (accuracy, speed, or prosody) [3, 4], producing global reading scores [5] or detecting specific errors [6, 7]. Variations in task and language [8, 5, 9] make it difficult to compare methods. Most ARA approaches use automatic speech recognition (ASR) system's transcriptions and acoustic features. A verbatim ASR would be desirable in this context. However, children's speech degrades ASR performance [10, 11] and systems trained on fluent speech often overlook disfluencies [12], a crucial information for assessment.

Some approaches for ARA rely on hybrid ASR systems designed to detect children's reading errors through customized decoding. Among them, task-specific decoding lattices (TSL) [13, 14, 8, 15] and n-gram-based language models (LMs) [3, 7] are commonly used. TSLs often rely on empirical weight tuning or limited-range hyperparameter search [16, 8], whereas n-gram LMs are typically trained on target prompts or annotated test data [7, 17]. For instance, [13, 14] use a two-layered approach where a phone lattice generated with an acoustic model and phone-level n-grams is decoded using a word lattice that models reading mistakes. Similarly, [8] introduces word-level lattices to capture corrections, false starts, and repetitions to

align children's speech. Other approaches use n-gram language models to constrain decoding. In [3] authors apply item-specific 3-gram LMs for real-time reading assessments on mobile devices, while [7] evaluates Kaldi-based systems with specialized 4-gram LMs to capture children's reading errors.

Recent work explores commercial and pre-trained ASR models for ARA. In [18], IBM Watson is evaluated in a clinical setting and found to be effective for basic classification but insufficient for fine-grained diagnosis. In [7], authors compare prompted and unprompted Whisper models, concluding that hybrid systems with specialized decoding strategies still perform better. The work in [6] examines end-to-end transformer-based phone-level decoders, noting that strong implicit language modeling can overlook reading mistakes. Finally, [19, 17] show that large self-supervised (SSL) models can be leveraged for ARA when fine-tuned on children's data to better handle reading errors and noisy conditions.

In this study, we evaluate pre-trained ASR systems for assessing reading fluency based on accuracy, speed, and reading error detection. We measure the system's ability to distinguish correct from incorrect words in read speech to predict the Correct Words Per Minute (CWPM) metric, a standard in reading assessment. A lower detection error rate results in better CWPM predictions. Additionally, we analyze ASR performance in terms of Word Error Rate (WER) and its capacity to detect disfluencies, using a Diagnosis Error (DE) metric.

Since the target texts are known and disfluencies follow predictable patterns, we leverage restricted decoding for improved recognition. We propose two approaches: (1) optimizing TSL arc weights via random search and (2) training n-gram LMs on synthetic texts enriched with disfluencies. Instead of using traditional phone-based ASR, we adopt an end-to-end character-level CTC model with finite-state transducer (FST) decoding for efficient lexicon and LM integration. Further, we explore in-domain fine-tuning.

Results indicate that integrating a controlled set of disfluency types into the decoding graph improves baseline performance and outperforms systems like Whisper[20]. The experiments are conducted on a dataset of Spanish-speaking children (ages 8–13) reading aloud in classroom settings under challenging noise conditions for ASR systems. Due to confidentiality constraints, audio samples cannot be shared. However, a key contribution of this work is the release of a text database with manually annotated transcriptions and labeled disfluencies (<https://zenodo.org/records/15522540>). Additionally, we share *disfluentES*, a toolkit for rule-based disfluent text generation for Spanish (<https://github.com/habla-liaa/disfluentES>).

Table 1: Average correct words per minute (CWPM), and percentage of words and sentences with disfluencies (%WWD and %SWD) over the complete dataset, and %WWD for each disfluency type on the subset of data annotated with this information.

Grade	All	3rd	5th	7th	8th
avg CWPM	114	77	108	133	128
%WWD	10%	14%	11%	9%	11%
%SWD	45%	47%	42%	44%	47%
%WWD per disfluency type					
SUB	2.97	3.69	2.80	2.41	2.98
PRE _{SUB}	1.35	1.76	1.56	0.74	1.33
PRE _{CUT}	1.32	1.78	1.41	0.77	1.29
REP	1.98	1.92	1.75	1.95	2.35
INS	0.85	0.83	0.79	0.76	1.03
DEL	0.50	0.49	0.33	0.56	0.66
PHO	0.28	0.56	0.26	0.08	0.20
CUT	0.04	0.09	0.03	0.02	0.04

2. Database

Our database is the outcome of a reading fluency census. It includes recordings of students from grades 3, 5 and 7 from primary school and 1st year of secondary school (hereafter referred to as 8th grade), ranging in age from 8 to 13. Each student read a level-appropriate text for one minute. The database contains 1327 speakers reading a total of 13043 sentences with 282, 309, 304 and 432 speakers from grades 3, 5, 7 and 8, respectively. The average word count per speaker is 86.72, 118.17, 144.21 and 141.61. The observed disfluency rates are shown in Table 1.

The recordings were collected in natural classroom environments using a wide range of mobile devices. This results in varied noise conditions, audio qualities and encoding formats. To assess speech quality (SIG), we use DNSMOS [21], an automated system that applies deep noise suppression based on Mean Opinion Score. Our analysis revealed significant differences in speech quality across grade levels (Kruskal-statistic=55.8, pvalue= $4.5e^{-12}$). To mitigate ASR performance degradation from noise, we apply FullSubNet [22] for audio enhancement, achieving a 27% relative improvement in SIG. However, significant differences between grades persist, with SIG ranking from worst to best in 8th, 3rd, 7th, and 5th.

Each recording has a verbatim orthographic transcription done by a linguist which captures disfluencies and pronunciation errors. Annotations excluded diacritics beyond standard accents, represented filled pauses with a single vowel and disregarded syllabification and vowel elongation, following pedagogical guidelines. Accents were retained only when mispronunciations formed real words. Given the near-transparent orthography of our target language, these transcriptions reliably reflect spoken content.

We annotated seven types of disfluencies for 450 speakers, following a modified version of Shriberg’s classification [23] as adapted in [8]. These include pre-corrections (PRE), where an initial attempt at a word is corrected; substitutions (SUB), involving word replacements or severe mispronunciations; phonetic errors (PHO), representing minor mispronunciations; repetitions (REP), verbatim re-readings of words from the original text; insertions (INS), added words not present in the text; deletions (DEL), omitted words; and cut words (CUT), where a word is truncated and left uncorrected. Table 1 provides detailed statistics for each category.

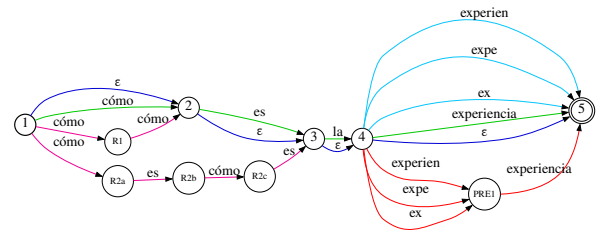


Figure 1: Illustration of the task-specific lattice (TSL) for the sentence “cómo es la experiencia”. The lattice accepts the original sentence (green paths) and cases with disfluencies. To simplify the example, we only show order-1 and order-2 REP of “cómo” and “cómo es” (magenta paths) and CUTS (cyan paths) and PRE_{CUT} (red path) for “experiencia”. DEL are accepted for all words (blue paths).

3. Disfluency generation

We develop *disfluentES* a toolkit for generating synthetic disfluencies from written text. We follow the methodology outlined in [24], using a rule-based approach and NLP techniques. The code, configurations and synthetic text dataset are publicly available in the repository provided above.

Our goal is to generate synthetic text that mirrors the disfluency patterns observed in our data. For this we can use the proportion of each type of disfluency (Table 1). Additionally, we identify the most affected POS tags per category. For the experiments presented here, we only used global statistics.

Specifically, we automatically annotate reference words with POS tags using SpaCy,¹ and introduce disfluencies from the seven types described above. For grammatical substitutions (SUB), we apply morphosyntactic modifications such as verb nominalization and changes in conjugation using mlconjug,² noun-adjective agreement and intra-class replacements for closed linguistic categories. For phonological errors (PHO) we rely on single-character edits, and for cuts (CUT) we eliminate final characters to obtain truncated variants. Further, we insert valid Spanish words (INS), duplicate the same word (REP), or apply pre-corrections (PRE), where an erroneous attempt precedes the correct candidate. These PRE are sub-classified into pre-correction cuts (PRE_{CUT}) and pre-correction substitutions (PRE_{SUB}), generated using the CUT and SUB heuristics. We also incorporate operation order to capture the complexity of multi-word transformations, such as duplicating two consecutive words in an order-2 repetition.

By applying these operations we generate a corpus with both correct and disfluent examples for each text. This corpus is then used to train an n-gram language model (LM) specific to that text, which should enhance the ability of the system to handle and recognize realistic reading errors without requiring text-specific recordings.

In addition to the disfluency generation approach above, we explore building task-specific lattices (TSLs) that contain a path that follows the reading prompt extended to include DEL, CUT, PRE_{CUT} and order-1 and order-2 REP disfluencies (Figure 1). The remaining disfluencies (INS, SUB, and PHO) are not implemented as arcs due to the challenge of generating realistic examples. This issue is further discussed in Section 6.

When using an n-gram model, the transcriptions are expected to reflect the statistics in the n-gram training data. In

¹<https://spacy.io/>

²<https://github.com/Ars-Linguistica/mlconjug3>

contrast, for TSLs, the transcriptions will be greatly affected by the arc weights which need to be determined empirically. Initially, we used the observed frequencies of each type of disfluency as proxies for arc weights, but this approach failed to produce reliable outcomes. To address this, we implemented a random search over the five arc weight corresponding to the disfluencies in the TSL. The values used for arc weights represent negative logarithm probabilities and were sampled uniformly between 0 and 20. This strategy provided a more flexible and optimizable approach to weight assignment.

4. ASR systems

For the purpose of transcribing speech with reading disfluencies, we use a traditional hybrid approach to ASR, integrating an acoustic model and a language model (LM). In end-to-end systems, the LM is learned directly from the data, making it challenging to apply targeted biases. In contrast, the hybrid approach allows for greater control over language constraints. Here, the acoustic model produces classification logits for a predefined set of tokens (e.g., phones or characters), which are then processed through a finite-state transducer (FST)-based decoding lattice to determine the most probable path.

For the acoustic model, we use a freely available character-level CTC Wav2Vec2.0 [25] ASR trained on Mexican Spanish data [26]. The model is the result of fine-tuning “wav2vec2-large-xlsr-53” for 5 epochs using approximately 944 hours of Spanish speech data collected or developed by the CIEMPIESS-UNAM Project since 2012.³ The model is trained in an end-to-end manner, but at inference, character-level logits are passed to a decoder. We employ an FST-based decoder implemented with the K2⁴ framework, constructed by composing three key FSTs: $\mathcal{H}$, in this case representing the CTC topology; $\mathcal{L}$, the lexicon; and $\mathcal{G}$, the grammar, which can be either an n-gram LM or a TSL derived from the reading prompt.

The n-gram LMs were built using the KenLM [27] toolkit with Kneser–Ney smoothing [28]. We experimented with different training sources, including manual transcriptions, reading prompts, and texts with synthetic disfluencies. Both 3-gram and 5-gram models were evaluated, with 3-grams performing as well as or better than 5-grams in most cases. Thus, the reported results are based on 3-gram models.

The TSLs were designed as described in Section 3. To determine the weights, we conduct a 100-iteration random search using cross-validation. The optimization criterion was the BER (see Section 5) applied separately for each grade.

In addition, we evaluate the effect of domain adaptation by fine-tuning (FT) the acoustic model with in-domain data. To do so, we apply a leave-one-grade-out cross-validation strategy, ensuring that each evaluation is performed on unseen prompts. We use 4.8h, 5.2h, 5.2h and 7.4h for 3dr, 5th, 7th and 8th, respectively. For comparison, we also include results from Whisper Large-v3.

5. Metrics

Human reading assessment involves evaluating different components of fluency and identifying specific reading errors. Correct Words Per Minute (CWPM) is one of the most widely used fluency metrics reflecting both reading speed and accuracy. For ARA, this metric is obtained from the ASR output which intro-

RP	cómo	-	es	la	experiencia	-	de	viajar	-
MT	cómo	e	es	la	esperancia	de	de	viaje	-
AT	cómo	e	es	-	espera	-	de	viajar	más
MT/RP	C	E	C	C	E		C	E	
AT/RP	C	E	C	E	E		C	C	
CM	TA	TR	TA	FR	TR		TA	FA	
COUNTS		CD			DE		DE		DE

Figure 2: Example of alignment between reading prompt (RP), manual transcription (MT) and automatic transcription (AT). For each pair, MT/RP and AT/RP, each slot (word or “-”) is labeled as correct (C) or error (E). These labels are then used to compute a confusion matrix containing the number of true acceptances (TA, for the pair CC), true rejections (TR, for EE), false acceptances (FA, for EC) and false rejections (FR, for CE). FR can be further categorized into correct diagnosis (CD) and diagnosis error (DE) to differentiate cases where the ASR correctly identifies the type of error from those where it fails to do so. ASR deletions (“-” in RP and AT but not in MT) and insertions (“-” in RP and MT but not in AT) are considered DE.

duces errors that may affect its estimated value. In this section, we describe the procedure used to evaluate the ASR output with respect to its usefulness to assess fluency.

Following [6], we use three sequences to evaluate our ARA systems based on ASR: Reading Prompts (RP), serving as the reference; Automatic Transcriptions (AT), produced by the ASR; and Manual Transcriptions (MT), the human-transcribed speech. ASR performance is measured by Word Error Rate (WER), calculated as the total number of insertions, deletions, and substitutions in AT relative to MT, divided by the total words in the MT. Lower WER indicates higher transcription accuracy, but a high WER does not necessarily imply a poor estimate of CWPM, since an ASR system may omit disfluencies yet still transcribe correctly-read target words without error.

To compute ARA metrics, RP, AT and MT sequences need to be aligned to each other. To do so we use a Multiple Sequence Alignment (MSA) method based on the center star heuristic [29], using RP as an anchor. Pairwise alignments use an adapted Needleman-Wunsch algorithm with a Levenshtein distance for scoring substitution candidates. This triple alignment defines the key cases for assessment (see Figure 2):

- True Acceptances (TA): Words are read correctly and are recognized as correct. Maximizing TA is crucial for accurate estimation of the CWPM since a high TA count indicates that the ASR correctly identifies well-read words, contributing to a precise score.
- False Rejections (FR): Words that are read correctly but are detected as incorrect. Minimizing FR is essential, as it leads to an underestimation of CWPM. A high number of FRs may cause frustration in children.
- False Acceptances (FA): Words that are read incorrectly but are detected as correct. Minimizing FA is important because these errors give credit to misread words, leading to an overestimation of CWPM. This may occur when the language model bias is strong, potentially causing the ASR to follow the prompt rather than accurately transcribing the audio.
- True Rejections (TR): Words that are read incorrectly and correctly detected as errors. While maximizing TR helps analyze children’s reading mistakes, it does not directly impact CWPM estimation. TR cases are divided into Correct Diagnosis (CD), when the ASR system correctly classifies the error type, and Diagnosis Error (DE) when it fails to do so. We

³<http://www.ciempiess.org/>

⁴<https://github.com/k2-fsa/k2>

Table 2: WER, BER and DER for different ASR systems across grades (3rd, 5th, 7th and 8th). The best-performing values for each metric (bold) are determined based on a 25% overlap in bootstrapped confidence intervals. FT stands for fine-tuned.

Level System	ASR WER				BER				DER			
	3rd	5th	7th	8th	3rd	5th	7th	8th	3rd	5th	7th	8th
Whisper Large-v3	21.5±2.6	14.1±2.1	13.8±1.9	19.3±2.2	13.3±2.3	7.4±1.7	8.5±1.6	13.7±2.3	36.3±6.5	33.0±6.1	29.2±6.1	40.0±5.2
Manual3gram	17.5±2.6	12.9±2.1	11.6±1.7	17.7±2.0	9.7±1.7	6.8±1.6	7.4±1.8	13.0±2.1	33.4±6.8	33.6±5.6	30.3±5.3	37.2±4.4
Prompt3gram	17.2±2.3	13.0±2.2	15.9±1.9	23.2±2.5	7.0±1.8	6.1±1.6	11.0±2.2	17.6±2.6	28.4±6.7	32.2±6.2	41.3±5.5	45.2±4.8
Prompt3gram-FT	17.6±2.8	12.8±2.3	16.1±2.2	22.3±2.4	8.0±1.9	6.4±1.7	11.9±2.1	16.9±2.2	28.8±6.3	36.1±5.9	41.9±5.1	46.1±4.4
RepDel3gram	17.3±2.5	13.0±2.1	12.1±1.6	18.2±2.1	7.3±1.9	5.8±1.5	6.5±1.4	11.5±2.0	27.5±6.8	29.3±6.5	35.2±6.0	42.1±5.7
RepDel3gram-FT	17.7±2.5	12.8±2.0	12.3±1.7	17.6±2.0	8.3±1.8	6.2±1.7	7.1±1.5	11.1±1.8	28.9±6.4	34.4±5.6	36.4±5.8	43.7±5.3
RepPreDelTSL	16.8±2.5	13.5±2.1	14.2±1.9	18.4±2.0	4.9±1.4	4.9±1.6	7.3±1.9	9.7±1.8	17.9±5.8	21.9±6.5	26.7±7.0	27.7±6.4
RepPreDelTSL-FT	17.0±2.5	14.3±2.3	14.9±2.2	18.8±1.8	5.7±1.3	5.9±1.8	8.3±2.1	10.8±1.8	19.1±6.7	27.7±7.1	30.4±7.0	28.3±6.3

define Diagnosis Error Rate (DER) as $DER = DE / (CD + DE)$.

- ASR deletions (“-” in RP and AT but not in MT) and insertions (“.” in RP and MT but not in AT) are classified as DE, as they lack a corresponding target word in the prompt, making rejection or acceptance irrelevant. This case differs from what is done in [6].

Our main goal in this work is to minimize the difference between the word count obtained from AT, given by $CW_{AT} = TA + FA$ and the ground truth word count, given by $CW_{MT} = TA + FR$. Yet, this difference is not a robust evaluation metric since it may be low even for very poor transcriptions, as long as $FR - FA \approx 0$. Hence, in this work we instead use the Balanced Error Rate (BER) metric, which is defined as $BER = FAR + FRR$. The False Acceptance Rate (FAR) measures how often a misread word is wrongly classified as correct and is given by $FAR = \frac{FA}{FA + TR}$. The False Rejection Rate (FRR) measures how often a correctly read word is wrongly classified as incorrect and is given by $FRR = \frac{FR}{FR + TA}$. Finally, ASR WER can be calculated from the confusion counts as $WER_{AT} = \frac{FR + DE + FA}{N}$.

6. Results and discussion

Table 2 shows the Word Error Rate (WER), Balanced Error Rate (BER) and Diagnosis Error Rate (DER) results for various ARA systems. We use two baseline systems: one based on an n-gram derived solely from the reading prompt (Prompt3gram) and Whisper Large-v3. Additionally, we consider the system that employs an n-gram obtained from manual transcriptions (Manual3gram) as a top-line reference. We apply a 5-fold cross-validation within each grade to avoid over optimistic results. Our goal is to approximate this top-line performance through task-specific language models (LMs).

Next, we introduce a system that incorporates n-grams learned from synthetic text containing disfluencies. We explored two approaches for generating this text: one based on simulating the population by considering the overall distribution of each disfluency type and another by generating a fixed number of occurrences per type. Additionally, we selected specific disfluency types to assess their impact. The best-performing system in this exploration, which used a fixed number ($n = 1000$) of REP and DEL, is included in Table 2 as RepDel3gram. For the WER metric, this system consistently matches or outperforms the others across all grades. In terms of BER, it performs as well as or better than both Manual3gram and Prompt3gram systems.

PRE, INS, and SUB did not improve performance for these models. PRE_{CUT} had a minor negative effect, while INS and SUB degraded results significantly. Understanding which types of SUB are beneficial for the task and are likely to appear in

new data is challenging as it affects the model’s ability to generalize effectively. Expanding the vocabulary with INS and SUB increases the number of candidate words, potentially raising the likelihood of misrecognitions. While n-grams with INS and SUB slightly improved FAR by preventing some miscorrections, they increased FRR, leading to more transcription errors.

The task-specific lattice (TSL) with disfluencies, referred to as RepPreDelTSL, exhibits less consistent WER performance compared to the 3-gram approach. However, it achieves the best BER results across most grades. Moreover, this system surpasses all others in terms of DER.

Inspecting the arc weights for the TSL after optimization, we found that the arcs with larger weights are, in increasing order, DEL, PRE_{CUT}, CUT, order-1 REP and order-2 REP. It is worth noticing that this ranking is different from that derived from the statistics of the corresponding disfluency types (Table 1), which is PRE_{CUT}, REP, DEL, and CUT. These results are in agreement with our initial observation that using the statistics to define the arc weights was highly suboptimal.

For the TSL, adding INS and SUB can be addressed by incorporating a lattice that accepts any sequence of tokens, similar to an open phone loop. However, this approach would require additional weight tuning while sacrificing the constraint imposed by the current method, potentially increasing variability and reducing overall robustness.

The effect of fine-tuning is inconclusive. While it led to improvements in certain settings (8th grade WER), it did not in others. This variability may be attributed to differences in acoustic conditions across grades, where the leave-one-grade-out approach could introduce mismatches. Future experiments should explore fine-tuning with cross-validation within individual grades to mitigate these inconsistencies.

Analyzing the behavior of the metrics with respect to speech quality (SIG), we found a consistent negative spearman correlation between FRR and SIG and between WER and SIG for all systems, with larger values for the Prompt3gram-FT ($\rho = -0.53, p \approx 0, \rho = -0.51, p \approx 0$). This confirms that ambient noise hurts ARA.

7. Conclusions

We integrated disfluency-aware decoding strategies into a Spanish ASR-based reading assessment system and evaluated it on our dataset of read children speech. By modeling common disfluencies, we decreased Word Error Rate (WER), Balanced Error Rate (BER), and Diagnosis Error Rate (DER). These findings confirm that explicitly accounting for disfluencies in the decoding process effectively captures children’s actual reading patterns under challenging acoustic conditions.

8. Acknowledgments

This work was partially supported by the World Bank. We acknowledge the predisposition of the Government of Mendoza for providing the audios they collected and collaborating with the ratings, and Vanesa Petran for annotating the audios.

9. References

- [1] L. S. Fuchs, D. Fuchs, M. K. Hosp, and J. R. Jenkins, “Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis,” in *The role of fluency in reading competence, assessment, and instruction*. Routledge, 2001, pp. 239–256.
- [2] V. M. Cabello González, “El estudio erce 2019 y los niveles de aprendizaje en ciencias. ¿qué nos dicen y cómo usarlos para mejorar los aprendizajes de los estudiantes?” 2022. [Online]. Available: <https://unesdoc.unesco.org/ark:/48223/pf0000382747>
- [3] J. Cheng, “Real-time scoring of an oral reading assessment on mobile devices,” in *Interspeech 2018*, 2018, pp. 1621–1625.
- [4] M. Duong, J. Mostow, and S. Sitaram, “Two methods for assessing oral reading prosody,” *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, pp. 1–22, 2011.
- [5] M. P. Black, J. Tepperman, and S. S. Narayanan, “Automatic prediction of children’s reading ability for high-level literacy assessment,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 1015–1028, 2010.
- [6] L. Gelin, M. Daniel, T. Pellegrini, and J. Pinquier, “Comparing phoneme recognition systems on the detection and diagnosis of reading mistakes for young children’s oral reading evaluation,” in *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 6–10.
- [7] B. Molenaar, C. Tejedor-Garcia, C. Cucchiari, and H. Strik, “Automatic assessment of oral reading accuracy for reading diagnostics,” in *Interspeech 2023*, 2023, pp. 5232–5236.
- [8] J. Proença, D. Celorico, S. Candeias, C. Lopes, and F. Perdigão, “Children’s reading aloud performance: a database and automatic detection of disfluencies,” in *Interspeech 2015*, 2015, pp. 1655–1659.
- [9] R. Gothi, R. Kumar, M. Pereira, N. Nayak, and P. Rao, “A dataset and two-pass system for reading miscue detection,” in *Interspeech 2024*, 2024, pp. 4014–4018.
- [10] S. Lee, A. Potamianos, and S. Narayanan, “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters,” *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.
- [11] A. Hämmäläinen, S. Candeias, H. Cho, H. Meinedo, A. Abad, T. Pellegrini, M. Tjalve, I. Trancoso, and M. S. Dias, “Correlating asr errors with developmental changes in speech production: A study of 3-10-year-old european portuguese children’s speech,” in *Workshop on Child Computer Interaction-WOCCI 2014*, 2014, pp. pp–1.
- [12] M. Zusag, L. Wagner, and B. Thallinger, “Crisperwhisper: Accurate timestamps on verbatim speech transcriptions,” in *Interspeech 2024*, 2024, pp. 1265–1269.
- [13] J. Duchateau, K. Demuyne, and H. Van hamme, “Evaluation of phone lattice based speech decoding,” in *Interspeech 2009*, 2009, pp. 1179–1182.
- [14] E. Yılmaz, J. Pelemans, and H. Van hamme, “Automatic assessment of children’s reading with the flavor decoding using a phone confusion model,” in *Interspeech 2014*, 2014, pp. 969–972.
- [15] M. Nicolao, M. Sanders, and T. Hain, “Improved acoustic modelling for automatic literacy assessment of children,” in *Interspeech 2018*, 2018, pp. 1666–1670.
- [16] T. Kouzelis, G. Paraskevopoulos, A. Katsamanis, and V. Katsouras, “Weakly-supervised forced alignment of disfluent speech using phoneme-level modeling,” in *Interspeech 2023*, 2023, pp. 1563–1567.
- [17] L. Gao, C. Tejedor-Garcia, H. Strik, and C. Cucchiari, “Reading miscue detection in primary school through automatic speech recognition,” in *Interspeech 2024*, 2024, pp. 5153–5157.
- [18] T. Piton, E. Hermann, A. Pasqualotto, M. Cohen, M. Magimai-Doss, and D. Bavelier, “Using commercial asr solutions to assess reading skills in children: A case report,” in *Interspeech 2023*, 2023, pp. 4573–4577.
- [19] L. Block Medin, T. Pellegrini, and L. Gelin, “Self-supervised models for phoneme recognition: Applications in children’s speech for reading learning,” in *Interspeech 2024*, 2024, pp. 5168–5172.
- [20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [21] C. K. Reddy, V. Gopal, and R. Cutler, “Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 886–890.
- [22] X. Hao, X. Su, R. Horaud, and X. Li, “Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.
- [23] E. E. Shriberg, “Preliminaries to a theory of speech disfluencies,” Ph.D. dissertation, Citeseer, 1994.
- [24] T. Passali, T. Mavropoulos, G. Tsoumakas, G. Meditskos, and S. Vrochidis, “Lard: Large-scale artificial disfluency generation,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 2327–2336.
- [25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [26] C. D. Hernandez Mena, “Acoustic model in spanish: wav2vec2-large-xlsr-53-spanish-ep5-944h.” 2022. [Online]. Available: <https://huggingface.co/carlosdanielhernandezmena/wav2vec2-large-xlsr-53-spanish-ep5-944h>
- [27] K. Heafield, “Kenlm: Faster and smaller language model queries,” in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.
- [28] H. Ney, U. Essen, and R. Kneser, “On structuring probabilistic dependencies in stochastic language modelling,” *Computer Speech & Language*, vol. 8, no. 1, pp. 1–38, 1994.
- [29] I. M. Wallace, G. Blackshields, and D. G. Higgins, “Multiple sequence alignments,” *Current opinion in structural biology*, vol. 15, no. 3, pp. 261–266, 2005.