10th Workshop on Speech and Language Technology in Education (SLaTE)
22-24 August 2025, Nijmegen, Netherlands

Refined Analysis of Reading Miscues

Gloria María Montoya Gómez[1], *Pol Ghesquière*[2], *Hugo Van hamme*[1]

[1]Department of Electrical Engineering (ESAT), KU Leuven, Belgium
[2]Faculty of Psychology and Educational Sciences, KU Leuven, Belgium
gmontoya@esat.kuleuven.be, pol.ghesquiere@kuleuven.be, hugo.vanhamme@kuleuven.be

Abstract

Miscue analysis offers valuable insights into reading development by examining oral reading errors, but is time-consuming and prone to human error, affecting early literacy instruction. This paper presents an automatic system combining an end-to-end Automatic Speech Recognition (ASR) model with a Weighted Finite-State Transducer (WFST) to align ASR output with expected pronunciations from target text. Trained on native Dutch-speaking children, it leverages multiple ASR hypotheses without relying on external forced-alignment tools. Results show effective reading assessment with false alarms under 5%, accurate counting of word attempts, and correct word matching in story reading. Miscue classification focusing on external phoneme detection shows promise results. New metrics are proposed to improve evaluation of attempt quality and miscue accuracy.

Index Terms: Reading Miscue Analysis, Automatic Reading Assessment, Children Speech Recognition, Weighted Final State Machines, Reading Error Detection.

1. Introduction

Learning to read involves decoding letters into sounds and recognizing whole words. Early readers rely more on decoding, while experienced readers depend more on word recognition. Although reading becomes faster over time, unfamiliar words remain challenging, and different types of errors, such as word misrecognition, can occur.

This study does not explore theories of how reading works in the brain but focuses on creating automated metrics with modern speech recognition to help researchers and educators understand learners' reading strategies, specifically during technical reading aloud. A common reading assessment metric counts words read correctly on the last attempt, but this ignores earlier attempts that can reveal insights into decoding versus recognition processes. This study aims to offer an Automatic Speech Recognition (ASR)-based approach to analyze all reading attempts.

First, the work analysis how accurately the number of reading attempts can be counted, a challenging task due to unclear or silent pronunciations. Secondly, it examines the relationship between reading attempts and the target word, emphasizing the significance of word beginnings and endings for recognition. This analysis uses a corpus with manual annotations that detail how each reading attempt corresponds to the start and end of the word.

This work also introduces a phoneme-based system for automatic reading assessment and miscue detection, trained and evaluated on oral reading data from Belgian Dutch-speaking children. The novelty of this work lies in: 1) Presenting a Weighted Finite State Transducer (WFST)-based approach to model reading tasks that enables efficient miscue analysis; 2) Developing a method to combine multiple ASR hypotheses that allows phonemes from different hypotheses to be considered as candidates at each time step, while also accommodating multiple possible output lengths; 3) Generating a diverse set of ASR hypotheses that scales well with utterance length is crucial, as simple beam search often overlooks important hypotheses in longer inputs; 4) Extending and refining miscue analyses to infer metrics about the reading attempts that are relevant to analyzing the learner's reading strategy; and 5) Proposing performance metrics for this analysis.

2. Methodology

The WFST-based miscue detection system proposed in [1] is extended to process multiple recognition hypotheses simultaneously. Compared to [2], the WFST-approach provides full control over the precision/recall trade-off in miscue detection. The workflow consists of: 1) Hypothesis generation (lower branch) 2) Parsing the ASR output to align each segment of the spoken input with its corresponding portion in the reference text; and 3) Assigning a reading strategy label to each attempt by comparing it with canonical pronunciation(s) of the word it was aligned to (see Fig. 1).

2.1. WFST-based error analysis

WFSTs are used as the core framework for modeling alignments and reading miscues, following standard formulations as described in [3]. The primary WFST operations used in this study are union, concatenation, and intersection. All modeling and calculations related to WFST were performed using the open-source Pynini library [4]. WFSTs are used for 1) Prompt modeling, representing canonical word pronunciation(s) of the words in the reading task, and 2) Hypothesis modeling, encoding multiple ASR phoneme hypotheses with weighted transitions to handle recognition uncertainty. The final output is obtained by performing the intersection of these two automata.

2.1.1. Prompt modeling

The proposed WFST models each word in the reading task based on its canonical phonetic form defined by a Dutch phonetic lexicon, covering all valid pronunciations. These word models are then concatenated from left to right to represent the complete task. Transition labels are based on the phonemes present in the datasets outlined in Section 3.1, collectively forming the vocabulary V with $|V| = 56$. The weights assigned to the arcs described in this section are found in Table 1. The key features enabling miscue detection in this work are:

10.21437/SLaTE.2025-41

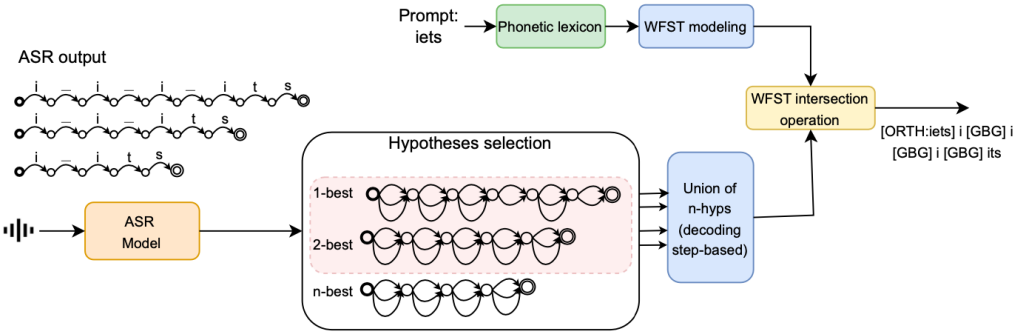

Figure 1: *Schematic overview of the automatic miscue detection model.*

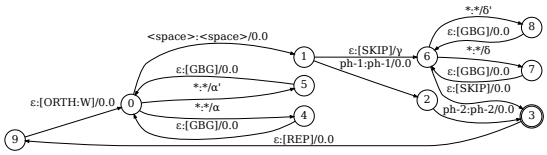

Figure 2: *WFST modeling of a single word when all the arc in Section 3.3 are implemented. The phonemes expected word the word "word1" are ph-1 and ph-2.*

1. ***Orthography arc***: Input ϵ-transition with the output label [ORTH:<w>], where <w> represents the upcoming word. This arc marks word boundaries, facilitating phoneme-to-word alignment. This transition has zero weight.
2. ***Garbage states***: Additional states (4 and 5) connect to the prompt through outgoing arcs α and α', which loop back via two input ϵ-transition labeled with [GBG] on the output. To model a word W consisting of n phonemes $p_1, p_2, \ldots, p_n$, arc α represents $|V| - n$ transitions, covering all phonemes in the vocabulary except those that make up the word by definition. Arc α' accounts for the remaining n transitions corresponding to the phonemes in W. This exclusion of phonemes is applied only at the word level and does not account for phonemes from adjacent words.
3. ***Skipping states***: This path consists of three states (6, 7, and 8). State 6 serves as a bridge between the initial state (following [ORTH]) and the word's final state(s). It's reached via an incoming arc γ, whose weight, acts as a scaling factor proportional to the number of phonemes in the word, as defined by the phonetic lexicon. The outgoing arc is an input ϵ-transition labeled with [SKIP] on the output. <space> is excluded from α to prevent consuming more than one word through this route.
4. ***Repetition***: Input ϵ-transition that links the final state back to the initial state, allowing word repetition after completion. This transition has [REP] as output label and zero weight.

2.1.2. Hypotheses modeling

The ASR model described in Section 3.3 uses beam search to generate several alternative hypotheses. However, it can still produce errors, especially with longer inputs, since maintaining enough hypothesis diversity would require increasing the beam depth significantly as the input length grows. This limitation may lead to false positives in reading error detection. The proposed hypothesis modeling is designed to address this issue, along with the drawbacks associated with the auto-regressive design of the ASR model [5]. To address these challenges, multiple alternative hypotheses of varying lengths are generated alongside the top prediction. Phonemes with low probabilities are excluded, and only those within the top 75% probability range are used, resulting in about 17 phonemes per decoding step for building the hypothesis WFST.

Building the WFST that encodes the information of the chosen n-hypotheses is created by taking the union of individual WFSTs for each hypothesis in the N-best list. The WFST of a single hypothesis is obtained by placing the surviving phoneme arcs in parallel and then concatenating over time, forming a directed acyclic graph (aka *sausage* network). Notice that a single hypothesis network has an exact score.

Table 1: *Weights applied in the WFST prompt modeling implementation. Weights marked with (*) indicate a scaling factor.*

Transition arc	Weight
γ^*	0.0009
$\alpha = \delta$	0.001
$\alpha' = \delta'$	0.0001

3. Experimental setup

3.1. Corpora and Lexica

The following corpora were used in stages, with their roles summarized in Table 2. Initial training was performed on adults' speech, followed by fine-tuning with children's speech.

- ***CGN:*** The Spoken Dutch Corpus (CGN) contains 900 hours of standard Dutch from Flanders and the Netherlands [6]. This work uses only the Flemish subset. A phonetic lexicon included in the corpus is used to convert orthographic transcriptions into phonetic annotations.
- ***CHOREC:*** The Children's Oral Reading Corpus (CHOREC) contains oral reading sessions from 400 Dutch-speaking children aged 6 to 12, with 92 hours of word- and phoneme-level annotations [7]. This dataset is used for training, development, and testing, with no speaker overlap between sets. The dataset includes isolated words and Story Reading Tasks (SRT) with varying levels of difficulty. The evaluation set

Table 2: *Data distribution for training. (*) represents values before data augmentation.*

	Adult's model		Children's model	
	Trn.	Dev.	Trn.	Dev.
Corpora	CGN	CGN	CHOREC JASMIN PVRT	CHOREC
Dur. (hrs)	150∗	1.5	82	4.5
#Samples	246750∗	2909	74463	4414
#Spks	1201	14	639	20

was taken from this dataset because of its rich human annotations, including miscue types, number of attempts, and final attempt classification.

Table 3: *Summary of key statistics from the evaluation set based on the reading strategy tier.*

Reading task	Dur. (hrs)	#Words	# Attempts	# Miscues
SRT-1	1.45	3037	4090	1053
SRT-2	0.76	2417	2834	417
SRT-3	0.76	2567	2870	303
SRT-4	0.85	3549	3886	337
SRT-5	0.88	4112	4449	337
SRT-6	1.38	6787	7355	568
SRT-7	1.58	8650	9285	635
SRT-8	1.90	11230	12032	802
SRT-9	2.53	17952	19016	1064

- ***PVRT:*** This *in-house* corpus includes 36 hours of Dutch readings by children and adults. Only Flemish speakers aged 5–12 were used for training. Phonetic transcriptions were derived from orthographic labels using a dictionary.
- ***JASMIN:*** JASMIN extends CGN including groups excluded from CGN such as children, non-native speakers, and elderly individuals from the Netherlands and Flanders [8]. This work uses only the subset of Flemish children aged 7–12. The corpus includes automatically generated phonetic labels.

3.2. Reading miscue labeling

As noted in Section 3.1, CHOREC was used as the evaluation set because of its fine-grained human annotations, organized into eight tiers. This study uses the tier named *reading strategies*, which includes 17 codes to represent reading miscues. This study focused on codes that provide information about the beginning, end, or entirety of a word: *ag/af* for correct/incorrect beginnings, *alg/alf* for correct/incorrect endings, and *g/f* for correct/incorrect full-word pronunciation. For example, if the word *bal* /bɑl/ (English: ball) is pronounced as /bɑ bɑl/, it would be assigned the codes *ag* and *g*. Based on the word-alignment by the WFST, each attempt was compared to the phonetic dictionary entry, and a code was assigned accordingly. After translating all miscues into the designated codes, a simple alignment was carried out. The evaluation verified whether both the code and its position corresponded to the reference. If they matched, the classification was considered correct; otherwise, it was counted as an error.

3.3. ASR System Parametrization

The speech-to-text model follows an encoder-decoder architecture using a hybrid CTC/Attention mechanism [9] trained on phoneme targets. It features a Conformer-based encoder [10] and a Transformer-based decoder [11]. The model's autoregressive design supports task-specific recognition but limits flexibility for unbiased automatic reading diagnosis [5].

The ASR system was trained using ESPnet2 in two stages: first on adult speech, then fine-tuned on children's speech using the pretrained encoder, improving recognition accuracy. In both models, the speech pre-processing involved extracting 80-dimensional Mel features, with SpecAug applied using ESPnet2 default setup (time warping, frequency, and time masking). The encoder and decoder consisted of 12 and 6 blocks, respectively, each with an embedding size of 256. The total number of trainable parameters was 30.61M. Training was performed with a batch size of 64 using the Adam optimizer.

The adult ASR model was trained for 100 epochs using phoneme targets, yielding a vocabulary of 50 units. Speed perturbation was applied with factors of $0.9, 1.0$, and 1.1 to triple the training data. The children's ASR model followed a similar setup, trained for 60 epochs with a 56-unit vocabulary to account for additional symbols in CHOREC's phonetic annotations that represent non-acoustic events. During inference, the CTC and decoder weights were set to 0.3 and 0.7, respectively.

3.4. Metrics

The accuracy of the ASR model is measured by using the metric Phoneme Error Rate (PER). This metric is chosen because it captures the subtle discrepancies between the output and the ASR ground truth. In this work, reading assessment classifies a word as correct if the final oral attempt matches the phonetic lexicon, based on Dutch standards [12]. Its performance is measured by False Alarm ($FA = F_p/(F_p + T_n)$) and Miscue Detection ($MD = T_P/(T_P + F_N)$), corresponding to false positive rate and recall. Miscue detection includes all earlier attempts, as multiple tries reflect deviations from fluent reading.

Let an evaluation set S consist of U utterances $u_1, u_2, \ldots, u_U$, where each utterance u_i is paired with an associated label Q_i. If u_i consists of N words, the label Q_i specifies the number of oral realizations per word, represented as $r_1, r_2, ..., r_N$. For an utterance u_i the system produces an alignment H based on its N target words. The count of attempts per each word of the alignment H is $h_1, h_2, \ldots, h_N$. Based on this, the *Attempt Detection Error Rate (AD.Err.)* is defined as

$$AD.Err = \frac{\sum_{u=1}^{U}\sum_{w=1}^{N}|h_w - r_w|}{\sum_{u=1}^{U}\sum_{w=1}^{N} r_w} \quad (1)$$

measuring how closely the automatic analysis matches the ground truth in terms of number of identified attempts. The *Attempt Class Error Rate (AC.Err)* compares the ground truth and inferred reading strategy codes. Recalling the evaluation set S defined before, and the aligned output obtained by the system H, if for each utterance u_i there are r miscues of which in k were correctly detected, the AC.Err is defined as

$$AC.Err = 1 - \frac{\sum_{u=i}^{U} k_i}{\sum_{u=1}^{U} r_i} \quad (2)$$

3.5. Weights fine-tuning protocol

Arc weights are crucial for obtaining a correct segmentation, balancing false alarms against missed miscue detections, and capturing the phonemes uttered rather than those expected from the target reading task. To ensure accuracy, focused subsets from the development set were created, each containing only one miscue type: hesitation and substitution (H+S), substitution (S), hesitation or false starts (H), and mixed cases (M). Results of the fine-tuning for each subset are presented in Table 4.

Table 4: *Performance per development subset. AC.Err is not reported for subset S, as it includes no words with more than one attempt.*

	Subset			
Metric	**H+S**	**S**	**H**	**M**
$Param.$	γ	α	α'	All
$Target$	FA	$AD.Err$	$AD.Err$	All
	MD	$AC.Err$	$AC.Err$	
$\%FA$	0.8	1.15	2.32	2.64
$\%MD$	91.3	88.1	66.6	88.4
$\%AD.Err$	9.1	4.46	3.37	10.8
$\%AC.Err$	22.7	-	21.88	34.3

4. Results

The ASR model trained exclusively on adults' speech achieved a PER of 24.2% on the SRTs. After fine-tuning with children's speech and applying data augmentation, the PER was reduced to 2.8% on the same dataset. The resulting accuracy is considered sufficient for both reading assessment and miscue analysis. A task-specific breakdown of the fine-tuned model's results is provided in Table 5.

Table 5: *Results reading assessment and miscue detection and classification using metrics described in 3.4*

SRT	**%PER**	**%FA**	**%MD**	**%AD.Err**	**%AC.Err**
1	7.2	4.9	93.8	14.6	45.8
2	3.5	2.9	86.3	7.9	32.9
3	3.5	3.1	88.3	7.9	27.3
4	1.9	1.6	82.1	7.9	30.1
5	1.6	2.1	78.6	5.6	34.9
6	2.2	2.9	90.1	5.5	28.8
7	2.6	4.3	80.2	5.9	31.7
8	2.3	3.4	77.3	5.9	32.1
9	2.9	4.2	81.7	5.5	35.6

A notable factor that negatively affects the performance of the ASR system, particularly in SRT1, is the handling of spacing between words. For example, if the intended word is *buiten* /bœtə{n}/ (English:outside) paired with the annotation /b œ bœ bœtən/ which shows three false starts separated by spaces /b œ bœ/, an ASR output /bœ bœ bœtən/ will be penalized for missing the spaces, which are interpreted as deletions.

Reading accuracy was found to be largely influenced by the number of hypotheses considered and the value of the weight γ. As noted in Section 2.1.2, incorporating a broad range of token options per hypothesis helps mitigate the limited output diversity typical of the long sequences, which characterize the evaluation set of this study. Consequently, careful tuning of γ was crucial to achieve this balance.

The *AD.Err* was found to depend on α and δ, which control phoneme assignment. Unequal weights can cause shorter words to absorb phonemes from neighbors, skewing miscue counts. Equal weights help reduce this bias. For example, consider the *een cameraploeg* /{ə|eː}n kaːm{ə|eː}rɑplux/ (English: a camera crew), where *een* is a short word in terms of phoneme count. If a speaker struggles with *cameraploeg*, those attempts should ideally be assigned to that word. However, if the weights of α and δ are not properly set, the system may find it less costly to assign the phonemes from these attempts to the shorter neighboring word, *een*.

5. Discussion

The results presented in this study are based on training the ASR model from scratch, integrating multiple hypotheses into a single lattice, and designing prompt arcs to account for potential miscues. A notable observation is the performance on SRT1, which shows the highest values for FA, AD.Err, and AC.Err. While higher error rates were anticipated due to the lower PER, another contributing factor is how word boundaries or blank spaces are handled. As shown in Table 3, SRT1 contains the highest average number of miscues per attempt, which aligns with the fact that it was read by the youngest group in the evaluation set. Early-stage readers often break words into parts when reading, and annotators mark pauses with spaces. However, since spacing is subjective, the model may merge separate miscues, leading to errors in detecting individual attempts.

The FA rate across the entire dataset remains below 5%, while the MD exceeds 80% in most cases. An exception to this trend is SRT5, which falls outside these ranges for both metrics. After a detailed review, it was identified that the model tends to recognize phonemes that result in a correct word, even if that differs from what was actually said.

To address the motivation of this work, both AD.Err and AC.Err must be considered together. The system effectively aligns false starts and hesitations with the correct parts of words. However, accurately counting attempts does not guarantee correct miscue classification, as the ground truth uses specific labels that the system must match. Errors can arise from how these labels are generated, not from the WFST itself. Overall, the results indicate that WFSTs are a promising approach for accurate miscue analysis.

6. Conclusions

This paper introduces a framework for reading assessment and miscue analysis in native Dutch-speaking children using an ASR model and a WFST that integrates ASR output, a phonetic lexicon, and the prompt to generate a reading error analysis. The system was evaluated on stories with varying reading difficulties.

The framework showed strong performance in reading assessment using the FA/MD ratio and accurately detected reading attempts per word. However, miscue classification was negatively impacted by alignment between the ground truth and inferred reading strategy codes. This paper proposed a method to automatically analyse reading attempts towards revealing the strategy learners use while reading a word. In future work, this information should be combined with word decoding/recognition times and error types (e.g. swaping letters).

7. Acknowledgments

The research was supported by KU Leuven Special Research Fund grant C24M/22/025 and the Flemish Government under the "Vlaams Artificiële Intelligentie Onderzoeksprogramma" program.

8. References

[1] G. M. Gomez, P. Ghesquière, and H. Van hamme, "Reading proficiency assessment using finite-state transducers," in *ICETC 2024*, 2024.

[2] L. Gao, C. Tejedor-Garcia, H. Strik, and C. Cucchiarini, "Reading miscue detection in primary school through automatic speech recognition," in *Interspeech 2024*, 2024, pp. 5153–5157.

[3] M. Mohri, F. Pereira, and M. Riley, *Speech Recognition with Weighted Finite-State Transducers*. Springer, 2008, ch. pp 559–584.

[4] K. Gorman, "Pynini: A python library for weighted finite-state grammar compilation," in *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, 2016.

[5] C. Cucchiarini, H. Strik, and L. Boves, "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms," in *Speech Communication*. Elsevier, 2000.

[6] N. Oostdijk, "The spoken dutch corpus. overview and first evaluation," in *Proceedings of the Second International Conference on Language Resources and Evaluation*, 2000.

[7] L. Cleuren, J. Duchateau, P. Ghesquière, and H. Van hamme, "Children's oral reading corpus (CHOREC): Description and assessment of annotator agreement." in *Proceedings of the International Conference on Language Resources and Evaluation*, 2008.

[8] C. Cucchiarini, H. Van hamme, O. van Herwijnen, and F. Smits, "JASMIN-CGN: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006.

[9] T. Hori, S. Watanabe, and J. Hershey, "Joint ctc/attention decoding for end-to-end speech recognition," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[10] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Computation and Language*, 2017.

[12] A. van Til, F. Kamphuis, J. Keuning, M. Gijsel, J. Vloedgraven, and A. de Wijs, *Wetenschappelijke verantwoording LVS-toetsen DMT*. Arnhem: Cito, 2018.