Reading Proficiency Assessment Using Finite-State Transducers

Gloria María, Montoya

Department of Electrical Engineering (ESAT), KU Leuven, gmontoya@esat.kuleuven.be

Pol, Ghesquière

Faculty of Psychology and Educational Sciences, KU Leuven, pol.ghesquiere@kuleuven.be

Hugo, Van hamme

Department of Electrical Engineering (ESAT), KU Leuven, hugo.vanhamme@kuleuven.be

This paper describes a model designed to detect miscues of children's oral reading. The model is evaluated on a corpus of primary school children with and without reading difficulties in Dutch. This set contains real words and pseudo words reading tests. The automatic speech recognition task is achieved by training an End-to-End (E2E) model with phonemic targets. The encoder employs the Conformer architecture, and the decoder follows the Transformer decoder scheme. The automatic assessment relies on modeling a lexicon at a phonetic level using a Weighted Finite State Transducer (WFST) that models the pronunciation lexicon. The proposed WFST construction accommodates all the pronunciations defined by the lexicon for any given word, allowing the assessment to handle multiple pronunciations. The experiments show that the accuracy of this setup outperforms previous results obtained on the same evaluation set.

CCS CONCEPTS • **Computing methodologies** → **Artificial intelligence** → **Natural language processing** → **Speech recognition**; • Applied computing → Education; • Theory of computation → Formal languages and automata theory → Automata extensions → Transducers.

Additional Keywords and Phrases: Reading assessment, Miscue detection, Pronunciation modeling, Reading tutor, Children speech recognition, Computer-assisted language learning, Finite state automata.

ACM Reference Format:

First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 INTRODUCTION

Recent advancements in Artificial Intelligence (AI), along with the availability of vast amounts of data and improved data processing technologies, have significantly enhanced the performance of Automatic Speech Recognition (ASR) systems. These developments have enabled ASR systems to achieve near-human levels of accuracy [1], leading to widespread use in applications such as real-time translation, customer service systems, and virtual assistants. Computer Assisted Language Learning (CALL) has also greatly benefited from

these ASR improvements, providing learners with precise, instant, and objective feedback essential for language proficiency.

Integrating CALL into children's education for reading and pronunciation offers several benefits. Poor early reading skills are a strong predictor of future academic challenges [2], and accurate pronunciation is crucial to avoid developing incorrect speech habits. A common metric for assessing reading skills is the number of words read correctly per minute (WCPM), typically evaluated during time-consuming one-on-one sessions with a human assessor. These evaluations can be inconsistent due to the assessor's expertise and subjective judgment. An automated tool for assessing reading proficiency would provide efficient, consistent, and scalable evaluations, saving time and resources while ensuring uniform assessments for all learners.

Developing accurate CALL tools for assessing children's oral reading is challenging due to the unique characteristics of children's voices and the complexities involved in ASR system development. Children's speech differs from adult speech in several ways: it has a higher pitch due to shorter vocal tracts and smaller vocal folds [3], slower and more variable speaking rates, and a developing vocabulary and pronunciation. Additionally, the maturation of children's vocal tracts and articulatory control varies with age [4]. Another major challenge is the limited availability of large, high-quality datasets of children's speech, which are crucial for training robust ASR systems. The scarcity of such data hampers the creation of models that can effectively generalize across different ages and speaking styles, making it difficult to address the dynamic variations in children's speech.

The aim of this paper is to introduce an automatic reading assessment system consisting of two components: 1) an E2E ASR system to convert the speech signal into a sequence of discrete phoneme symbols, and 2) a WFST that leverages a pronunciation lexicon to formally define what is considered a miscue, while allowing for multiple pronunciations. The system is evaluated on Belgian Dutch (Flemish) native children completing single word reading tasks, outperforming the previous best results published in [5]. To address issues previously mentioned, specifically data scarcity, the system exploits data augmentation strategies as well as fine-tuning adults' speech system fine-tuning to enhance the performance of the ASR system. The performance of the pronunciation assessment is compared in task-dependent (TD) and task-independent (TI) scenarios, meaning evaluating models on tasks they were and were not specifically trained, respectively. Although oral tests are typically known in advance, testing the TI case is particularly interesting as it simulates a future situation where new assessments are introduced without updating the ASR component.

The paper is organized as follows: Section 2 provides a summary of the fundamental concepts upon which the main components of the framework were built, section 3 describes the data used, the parametrization of each component of the system and the metrics employed in the experiments, section 4 present the results from testing various configurations, compares the performance of TD against TI, and provides the receiver operating characteristics curves for each task. Lastly, section 5 draws some conclusion and future experiments to perform.

Research question: When applying E2E ASR technology to reading assessment, several design choices must be considered. The investigation focuses on the following questions: 1) Is there a benefit in pretraining the acoustic model with a larger quantity of adult speech in addition to the (limited) child speech? 2) Does the trained decoder maintain task dependence despite being trained on a mixture of in-domain and out-of-domain data? 3) What is the optimal miscue detection criterion to use in the WFST?

2 ARCHITECTURE

The workflow of the proposed system involves two main steps: 1) converting the acoustic speech signal into text; 2), comparing the resulting text the with expected pronunciation(s) given a reading task. This process is depicted in Figure 1. The speech recognition is executed by a model trained from scratch, while the pronunciation analysis employs WFSTs. This work adheres to phoneme modeling because this provides an analysis related to correct perception of the words read, thereby facilitating accurate pronunciation assessment.

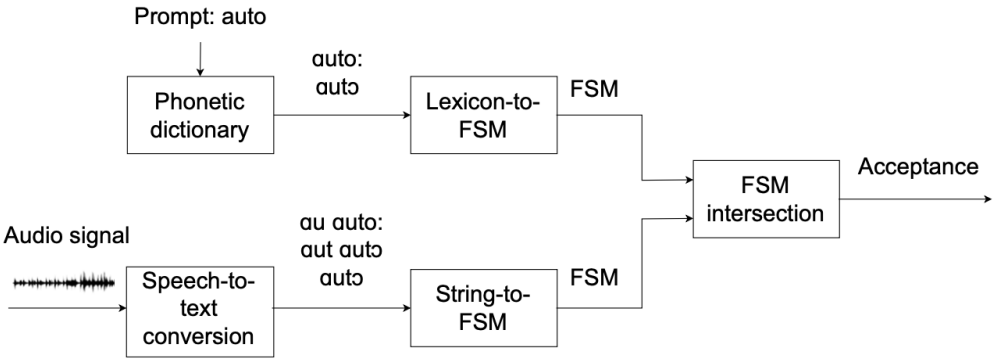


Figure 1. Workflow and components proposed for the automatic pronunciation system. Example for prompt “auto” (car) with pronunciation alternatives in IPA (top) and ASR hypotheses (bottom).

2.1 Speech-to-text Model

The ASR model is based on the encoder-decoder framework using a hybrid CTC/Attention [6] learning approach. The encoder follows the Conformer architecture [7], while the decoder adapts the Transformer architecture [8]. The encoder effectively models local and global contexts of the acoustic input, while the decoder captures long-range dependencies, meaning that can generate context across extended speech segments. For detailed information on each component of these architectures, please consult their respective papers. The ASR’s final output consists of a list of recognized phoneme strings, ranked in descending order of their cumulative probability.

2.2 Weighted Finite State Transducers

A Weighted Finite-State Transducer (WFST) is a mathematical abstract model that can be defined as a tuple $G = (Q, \bar{q}, \Sigma, M, \delta, Q_f)$, where Q is the set of states, $\bar{q} \in Q$ is the initial state, Σ is the set of input symbols (phonemes here), M is the set of output symbols, $\delta \subseteq Q \times \Sigma \times M \times \mathbb{R} \times Q$ is the set of transitions, and $Q_f \subseteq Q$ is the set of final states [9]. A transition $t_i = (q_{i-1}, \sigma_i, \mu_i, \nu_i, q_i)$ from state q_{i-1} to q_i consumes input symbol σ_i at a cost of ν_i, while emitting output symbol μ_i. An input string $\omega = \sigma_1 \dots \sigma_l$ is accepted by G if there exist transitions $t = (t_1 \dots t_l)$ such that $q_0 = \bar{q}$ and $q_l \in Q_f$. The total weight associated with the input string is computed using the weight aggregation function $R(\nu_1, \dots, \nu_l)$. For research question 3, the following functions are considered: strategy 1 $R_1(\nu_1, \dots, \nu_l) = \sum_{i=1}^{l} \nu_i$, strategy 2: $R_2(\nu_1, \dots, \nu_l) = \frac{1}{l}\sum_{i=1}^{l} \nu_i$, strategy 3: $R_3(\nu_1, \dots, \nu_l) = \sum_{i=k+1}^{l} \nu_i$, strategy 4:$R_4(\nu_1, \dots, \nu_l) = \frac{1}{l-k}\sum_{i=k+1}^{l} \nu_i$ where k is the position of the last [STUT] (cfr. infra) output symbol. Strategies 3 and 4 score only the quality of the last reading attempt, rather than including all attempted miscues.

In this paper, each phoneme symbol in an ASR hypothesis (section 2.1) has an associated weight $w_i = \log P(\sigma_i | \sigma_1, \ldots, \sigma_{i-1}, X)$ where X is the input speech. This weight is added to v_i in the aggregation function when σ_i is consumed. Finally, the n (design parameter) ASR hypotheses with the highest accumulated w_i values are offered to the WFST, and the aggregation function is computed for the hypotheses it accepts. A reading attempt is considered correct when any hypothesis leads to an aggregation score above a threshold β. This threshold determines the trade-off between the number of falsely detected reading errors and the number of missed reading errors.

3 EXPERIMENTS

3.1 Speech Data

3.1.1 CHOREC

The Children's Oral Reading Corpus (CHOREC) [10] is a dataset that contains reading sessions from Dutch-speaking children aged 6-12, enrolled in regular or special education schools. The speech data was collected from 8 different schools (S01 – S08). Further details of this dataset can be found in Table 1. The reading tasks are categorized into two types: isolated words and stories. The isolated word reading task comprises individual words varying in length and meaning. These words range from one to four syllables. In terms of meaning, they can be real words or pseudo words—orthographies that adhere to Dutch phonology but do not form existing words. Each sample is paired with its corresponding word and phoneme-level labels, along with descriptions of reading errors, reading strategies, and background noise. All these labels were manually annotated and serve as the ground truth or target output.

Table 1. Summary of the datasets used for training and evaluation.

Feature	CHOREC	PRVT	Jasmin	CGN
# Samples	94404	25334	9604	361535
# Speakers	399	411	48	1181
Male	234	218	23	707
Female	165	193	25	474
Duration	91h55m	36h16m	3h14m	151h41m

3.1.2 PRVT

The PRVT is an "in-house" database that contains recordings of children and adults reading isolated words and sentences in Dutch. This paper exclusively utilizes samples from Flemish native speakers aged 5 to 12. Detailed information about this subset is provided in Table 1. The dataset features orthographic annotations, except for mispronounced words, which are transcribed at the phoneme level. Word-level labels were converted via dictionary lookup.

3.1.3 JASMIN

JASMIN [11] contains contemporary Dutch spoken by children, non-native speakers with diverse mother tongues, and elderly individuals from the Netherlands and Flanders. This work focuses only samples from native Dutch (Flemish) children age from 7 to 12 who performed the reading task. Table 1 shows detailed information

about this subset. JASMIN's labels were automatically generated by forced alignment based on lexical lookup of the orthography. No conflict between CHOREC and JASMIN phoneme sets was encountered.

3.1.4 CGN

The Spoken Dutch Corpus (CGN) is a corpus that contains approximately 900 hours of contemporary standard Dutch as spoken in Flanders and the Netherlands [12]. The entire corpus is transcribed both orthographically and phonetically, with the latter generated automatically. CGN is divided into several components, with the two most distinct groups being dialogue/multilogue and monologues. Comprehensive information about these groups can be found in [12]. This paper exclusively utilized data collected from Flemish speakers. The most relevant features of the data used from CGN is shown in Table 1. The total number of speakers by gender is an approximation, as some samples are labeled with unknown identities.

3.2 ASR system parametrization

3.2.1 ASR model

All the ASR experiments were conducted using the ESPnet toolkit [13]. The architecture described in Section 2.1 consists of 12 Conformer encoder layers and 6 Transformer decoder layers, all with an embedding size of 256. For detailed information of the training setup, refer to the configuration files (The ESPnet configuration files will be made available on github). The phonemes were tokenized with Byte Pair Encoding (BPE) subword segmentation, resulting in a vocabulary of 300 units. No external language model was used, as it could mask mistakes by providing contextual information and lead the system to reconstruct existing words, similarly to the decoder's behavior mentioned in Section 3.1. The model should precisely detect what was uttered, regardless of whether the resulting word is linguistically valid.

3.2.2 Data distribution

The development of the child speech ASR model, as outlined in Section 2.1 involves two distinct stages: 1) training an ASR model from scratch using adults' speech with phoneme targets, and 2) using the encoder from step 1 to initialize and train the child speech ASR model. Details of the training and validation sets used for developing the adult ASR system are provided in Table 2. Speed Perturbation (SP) [14], a speech recognition augmentation technique, was applied using factors of 0.9 and 1.1 with the Sox audio manipulation tool. SP was applied to the training set for both adult and child models, resulting in a 3-fold expansion of the training data each. Additional information on the data distribution used for training both models are provided in Table 2 and Figure 2.

Table 2. Data distribution used for the training of the adults and children ASR models (TD)

	Adult's model		Children's model	
Feature	Training	Validation	Training	Validation
Dataset	CGN	CGN	CHOREC, PRVT, Jasmin	CHOREC
# Samples	358787	2748	82071	5192
# Speakers	1166	15	639	20
Male	699	8	363	13
Female	467	7	276	7
Duration	149h54m	1h47m	82h6m	4h55m

	Adult's model		Children's model	
Feature	Training	Validation	Training	Validation
Augmentation dur.	302h50m	-	164h7m	-

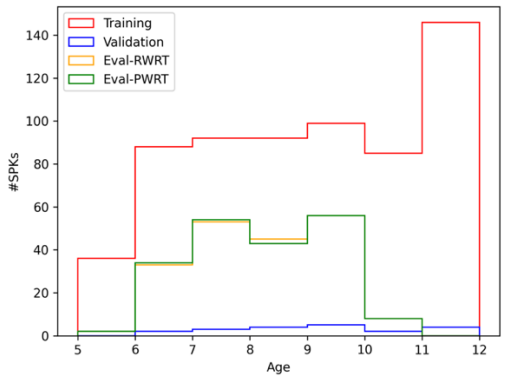


Figure 2. Distribution of children's age across the training, validation and evaluation subsets

The evaluation set consists of 199 speakers from CHOREC, spanning from schools S01 to S03. Further information of the different evaluation sets, Real Word Reading Task (RWRT) and Pseudo Word Reading Task (PWRT), is provided in Table 3 and Figure 2.

Table 3. Distribution of evaluation set based on the reading tasks

	RWRT			PWRT		
Feature	1-Syl	2-Syl	3+4-Syl	1-Syl	2-Syl	3+4-Syl
# Samples	7684	7135	6282	7793	7067	6118
# Speakers	191	182	160	196	178	154
Male	96	92	82	97	90	78
Female	98	90	78	99	88	76
Duration	4h14m	4h13m	4h3m	5h29m	5h32m	6h3m

The duration of the training data for the TI case was reduced by 30 hours, resulting in a total duration of 134 hours of children data. The reading tasks 2LG and 2LGPseudo were held out during the training process.

3.3 WFST parametrization

WFSTs were employed to model both the pronunciation lexicon, and the output generated by the ASR system. The library utilized for both scenarios was Pynini [15], an open-source Python library designed for finite-state text processing. The orthography of the reading task is converted into a WFST with Σ equal to the 45-element phoneme set (top branch of Figure 1). The WFST architecture proposed in this work facilitates the inclusion of multiple pronunciations per word and accommodates errors at any position within the word. These capabilities were achieved in the isolated word scenario using WFSTs executing the following steps:

1. For each valid pronunciation of a word, an individual WFST is created with all $\mu_i = \sigma_i$ and $\nu_i = 0$ arc weights set to 0. The union of these WFSTs is then performed, and redundant states are eliminated through optimization.
2. Addition of returning arcs with $\sigma_i =< \varepsilon >$, $\mu_i = [STUT]$ and $\nu_i = 0$ from each state back to the initial state in the resulting WFST from step 1. Here <ε> indicates the transition can be taken without consuming an

input symbol. These arcs enable the model to restart the assessment if stuttering or mispronunciation is detected, which is crucial because the reading assessment requires the reader to fully pronounced the prompted word.

3. Inclusion of a garbage state to capture false starts or unrelated words spoken during the reading task. This state is incorporated through transition arcs that originate from and point back to the initial state. The arcs connecting from the initial state are assigned a weight of 1 to penalize undesirable speech. The outgoing arc to the initial state was given the input label <ε> indicating that it can loop back to the beginning without consuming an input symbol.
4. Insertion of a state between the initial state and the first state corresponding to the expected input, creating a clear distinction between false starts and a successful pronunciation attempt. The input label assigned to this arc is *<space>* because it is desired to consider a reading as correct if an instance of the word itself was sounded out instead rather than sub parts. For example, based on Figure 1, "sɑuto:" would be rejected, while "s ɑuto:" would be labeled as correct. Every ASR hypothesis is proceeded by a <space>.

The resulting WFST, after applying the above aspects in the model's structure, can assess whether an utterance is correctly pronounced or if substitutions occur in the spoken word. Figure 3 depicts the automata after these adjustments. Classification of substitution types is not addressed in this study.

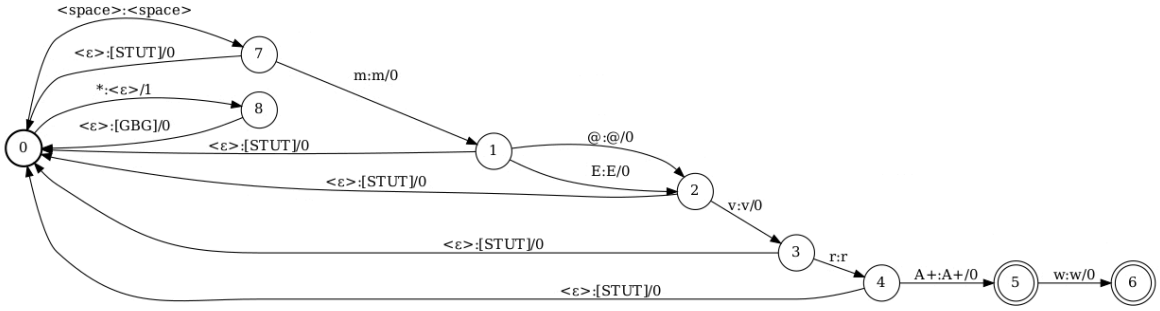


Figure 3. WFST of target "mevrouw" (pronunciation: m@vrA+ /məvrɑu/, m@vrA+w /məvrɑuw/, mEvrA+w /mɛvrɑuw/ or mEvrA+ /mɛvrɑu/). Resulting graph after following the steps described in section 3.3. The "*" accepts any $\sigma \in \Sigma$. The edges are marked as σ_i: μ_i/ν_i.

3.4 Metrics

As the system described in this paper performs binary classification to determine whether an utterance is correctly pronounced, the evaluation metrics used are derived from the confusion matrix. The metrics are: 1) False Alarm rate ($FA = Fp/n$) is the fraction of correctly read words flagged as incorrect, and 2) Miscue Detection rate ($MD = Tp/p$) understood as the proportion of the reading errors that was classified correctly as a reading error, i.e., recall. In this context, p=positive, n=negative, T=true, and F=false. These metrics were also used in studies that developed models for the same task, enabling direct comparison between systems [5].

4 RESULTS

The ASR model constitutes the first step of the proposed framework, making the verification of its accuracy essential, as its output is used in the subsequent stage. The overall Character Error Rate (CER) for the model trained exclusively on children's data was 6.2%. However, by incorporating adult data and applying SP, the CER improved to 5%, with the lowest and highest values being 7.3% for 1-syl PWRT, and 2.8% for the RWRT, respectively. This answers the first research question positively. The ASR task was considered sufficiently accurate for conducting several reading assessment experiments, under the assumption that any discrepancies between the ground truth and the WFST output were attributable to the architecture detailed in section 3.3.

Relying exclusively on the 1-best hypothesis from the decoder to perform the assessment would restrict the system's performance by confining the WFST to a single possible path, thus offering limited error handling capabilities. Flexibility between FA and MD rates is key because it ensures that the feedback is both accurate and constructive, preventing unnecessary confusion for learners.

Various methods for effectively integrating multiple hypotheses from the decoder along with their log probabilities were tested. The factors to adequately incorporate this data are 1) effect of setting an acceptance threshold, 2) optimal number of hypotheses to include in the intersection and how this number affects the results, and 3) how to utilize the probabilities per vocabulary unit when handling long output strings.

The results to be discussed are illustrated in the various images of Figure 4. Image 1 illustrates the tradeoff the between MD and FA rates as the threshold β varies (dotted red line) using strategy 1. It is observed that as the threshold increases, FA decreases. This behavior was expected, as a higher threshold allows the system to accept a broader range of confidence scores. The same image also demonstrates that FA and MD rates stabilize after a certain number of hypotheses are used. Multiple experiments across different reading tasks revealed that using n>20 yields steady results; therefore, obtaining more than 20 hypotheses from the decoder does not offer additional tradeoff between these rates. Image 2 displays the effects on how the weights are used. Longer hypotheses tend to have higher cumulative probabilities due to the increased number of BPEs, making them less likely to be selected by the system, which prioritizes paths with lower cost. Results show that normalizing the probabilities in combination with $w[t_i]$ in combination with a restriction of i yields to the lowest FA.

Figure 5 presents the results for all the reading tasks using the most optimal setup tested, which involves limiting the input string to its final attempt applying weight normalization. This configuration answers research question 3 as "strategy 2: $R_2(v_1, \dots, v_l) = \frac{1}{l}\sum_{i=1}^{l} v_i$". Also note that the proposed system significantly outperforms the previous best results published in [5] on this evaluation set. While the former system achieved 47% MD rate with 3% FA rate, the system presented in this paper achieves 77% MD rate at 3% FA rate.

The final analysis of this paper examines the system's performance in the TI case. Figure 6 shows that in the TI case, even with an increased threshold, the FA rate remains lower than the TD case. Despite the large amount of training data that is unrelated to the task, the decoder still seems to learn the mistake patterns leading to a significant advantage in TD performance over TI performance, which also answers research question 2. Nevertheless, the TI demonstrate that the system still provides consistent performance, as its FA rate remains lower than the Children Error rate (Ch. Err.), i.e., the overall percentage of reading errors made by children.

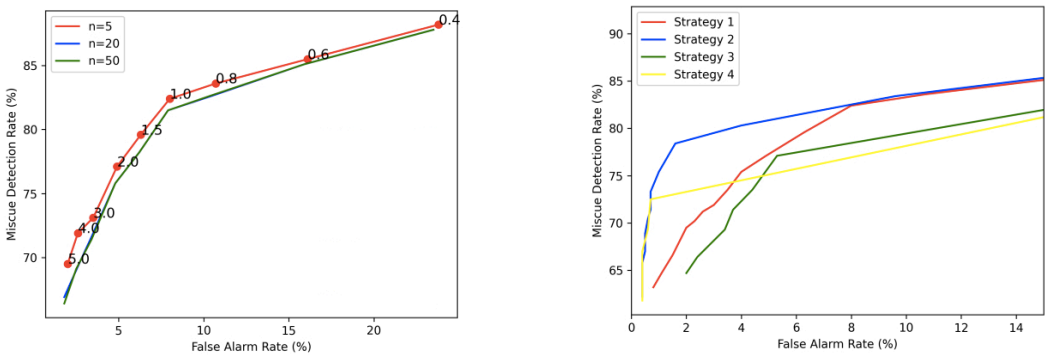


Figure 4. Image 1 Analysis of the impact of increasing the threshold and the number of hypotheses from the decoder. Image 2 Analysis of changing the acceptance criteria according to Section 3.4. Results were obtained from 2-syl RWRT.

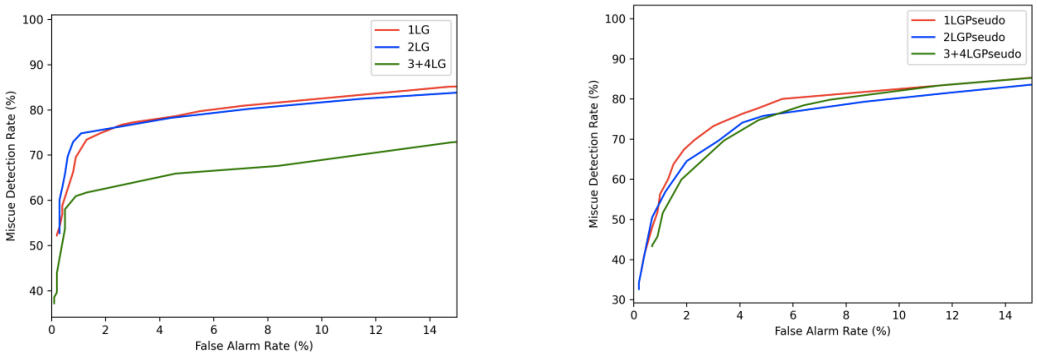


Figure 5. Results of each reading task individually. Image 1 Results for RWRT. Image 2 Results the results for PWRT

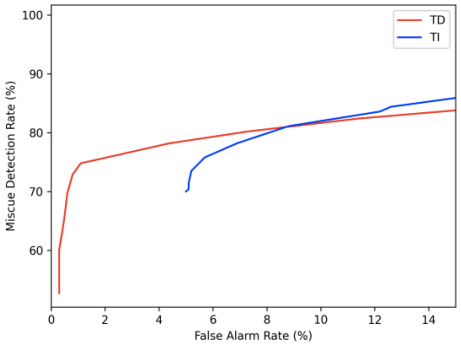


Figure 6. Results of TD and TI on 2LG. Ch. Err: 7%

5 CONCLUSIONS AND FUTURE WORK

In this paper, a framework was proposed for assessing reading accuracy of children's speech. The model consisting of a phoneme-based E2E recognizer with a WFST utterance model sets a new standard on the

CHOREC reading evaluation in Dutch for all performance metrics defined earlier for the task. A method for controlling the FA/MD trade-off was proposed. The actual operating point depends on the intended application. For the evaluation of the reading proficiency, a working point with a higher FA and MD rate can be chosen, especially for the less proficient readers or for the pseudo-word reading tasks, where more reading miscues are expected. For applications where corrective feedback is given, lower FA setting would be preferred, especially for the more proficient readers. The FA rate can be adjusted well below 1%, while still catching around half of the reading errors. Encouraged by these positive results, we see several directions for future work: (1) Refining the method to represent the hypotheses retrieved from the ASR system (2) Leveraging adult and multilingual children's data (3) Refining the WFST to detect and classify specific reading miscues (4) Designing corrective feedback strategies tailored to the performance capabilities of the reading miscue detection (5) Extend the assessment architecture to support sentences.

ACKNOWLEDGMENTS

This research was supported by KU Leuven Special Research Fund grant C24M/22/025 and the Flemish Government under "Onderzoeksprogramma AI Vlaanderen".

REFERENCES

[1] C. Spille, B. Kollmeier and B. T. Meyer, "Comparing human and automatic speech recognition in simple and complex acoustic scenes," *Computer Speech & Language,* vol. 52, pp. 123-140, 2018.

[2] J. Torgesen, "Catch Them Before They Fall: The Psychoanalysis of Breakdown," *The American Educator,* vol. 22, pp. 32-39, 1998.

[3] S. Lee, A. Potamianos and S. Narayanan, "Analysis of Children's Speech: Duration, Pitch and Formats," in *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1997.

[4] S. Das, M. Picheny and D. Nix, "Improvements in children's speech recognition performance," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98,* vol. 1, pp. 433-436, 1998.

[5] E. Yilmaz, P. Joris and H. Van hamme, "Automatic Assessment of Children's Reading with the FLaVoR Decoding Using a Phone Confusion Model," in *Proc. Interspeech 2014*, Singapore, 2014.

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

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

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention Is All You Need," in *Neural Information Processing Systems*, California, 2017.

[9] M. Mohri, F. Pereira and M. Riley, "Weighted finite-state transducers in speech recognition,," *Computer Speech Language,* vol. 16, pp. 69-88, 2002.

[10] L. Cleuren, J. Duchateau, P. Ghesquiere and H. Van hamme, "Children's Oral Reading Corpus (CHOREC): Description and Assessment of Annotator Agreement," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, 2008.

[11] 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 (LREC'06)*, Genova, 2006.

[12] N. Oostdijk, "The Spoken Dutch Corpus. Overview and First Evaluation.," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, 2000.

[13] S. Watanabe, H. Takaaki, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Yalta Soplin, J. Heymann, M. Wiesner, C. Nanxin, R. Adithya and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech 2018*, Hyderabad, 2018.

[14] T. Ko, V. Peddinti, D. Povey and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *Proc. Interspeech 2015*, Dresden, 2015.

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