Evaluation of Reading Performance of Primary School Children: Objective Measurements vs. Subjective Ratings

Erika Godde[1], *Gérard Bailly*[1], *David Escudero*[2,1], *Marie-Line Bosse*[3] *and Estelle Gillet-Perret*[4]

[1] GIPSA-Lab, Univ. Grenoble Alpes & CNRS, F-38000 Grenoble, France
[2] Department of Computer Science, Univ. of Valladolid, Spain
[3] LPNC, Univ. Grenoble Alpes & CNRS, F-38000 Grenoble, France
[4] CHU de Grenoble, Centre du Langage, Grenoble, France

{erika.godde,gerard.bailly}@gipsa-lab.fr, descuder@infor.uva.es,
marie-line.bosse@univ-grenoble-alpes.fr, EGilletperret@chu-grenoble.fr

Abstract

We analyze here readings of the same reference text by 116 children. We show that several factors strongly impact subjective rating of fluency, notably number of correct words, repetitions, errors, syllables spelled per minute. We succeeded in predicting four subjective scores – rated between 1 and 4 by human raters – from such objective measurements with a rather high precision ($R > .8$ for 3 out of 4 scores). This open the way for automatic multidimensional assessment of reading fluency using calibrated texts.

Index Terms: reading, pauses, evaluation, fluency, children

1. Introduction

Most literacy educators consider fluency to be a critical component of reading development [1, 2]. Although there are a number of definitions of reading fluency, there is a large consensus that accuracy, automaticity, and prosody all make an important contribution to the – listener-oriented – reading quality and – speaker-oriented – text comprehension. Kuhn et al. [3] define fluency as:

> "[...] combines accuracy, automaticity, and oral reading prosody, which, taken together, facilitate the readers construction of meaning. It is demonstrated during oral reading through ease of word recognition, appropriate pacing, phrasing, and intonation. It is a factor in both oral and silent reading that can limit or support comprehension."

Most oral fluency scales typically distinguish between four major steps of reading development that clear impact on perceived reading fluency (see Table 1):

Word processing that mainly consists in successfully accessing adequate lexical entries (pronunciation and semantics) without identifying grapho-phonetic constituents (phones, syllables)

Grouping that mainly consists in successfully on-line coordinating identification of words and speech planning, i.e. crossing the word barrier.

Phrasing that mainly consists in successfully pacing word grouping into meaningful linguistic units.

Expressivity that mainly consists in successfully computing adequate prosodic patterns from the on-line comprehension of the content.

Reading automaticity is considered to be acquired after the two first stages: it's the symptom of the successful synchronization (Breznitz [4]) of the brain entities involved in the visual, orthographic, phonological, semantic processing of textual input and the planning and on-line control of speech output. Our work observes pupils when achieving this skill, that is supposed to happen within the primary school, i.e. between 9 to 10 years.

Table 1: *Oral fluency scale (from Dpt of Education. NAEP, 2002 Oral Reading Study).*

Fluent	Level 4	Reads primarily in larger, meaningful phrase groups. Although some regressions, repetitions, and deviations from text may be present, these do not appear to detract from the overall structure of the story. Preservation of the authors syntax is consistent. Some or most of the story is read with expressive interpretation.
	Level 3	Reads primarily in three- or four-word phrase groups. Some small groupings may be present. However, the majority of phrasing seems appropriate and preserves the syntax of the author. Little or no expressive interpretation is present.
Nonfluent	Level 2	Reads primarily in two-word phrases with some three- or four-word groupings. Some word-by-word reading may be present. Word groupings may seem awkward and unrelated to larger context of sentence or passage.
	Level 1	Reads primarily word-by-word. Occasional two-word or three-word phrases may occur - but these are infrequent and/or they do not preserve meaningful syntax.

2. State of the art

2.1. Objective assessment of fluency

2.1.1. L2 learners

Most works in state of the art focusing on automatic evaluation of fluency belong to the field of computer assessment pronunciation training (CAPT). In this domain, main activity has been oriented to the evaluation of the segmental quality of L2 learners [5] and less to its prosodic quality. There exist commercial systems for training L2 oral proficiency (see table 2 of [6] for a comprehensive list of commercial systems) most of them, proposing reading activities. Nevertheless, oral reading is not the main concern for testing language proficiency (with tests like TOEFL® iBT [7] and others), because *speaking* competences are evaluated from spontaneous speech or oral discourses (describing pictures or expressing opinions) and *reading* competences are evaluated with questions about the understanding of a proposed text.

Concerning the evaluation of fluency, prosody has been the main concern of a number of works in the state of the art of L2 pronunciation analysis (see [8] for an updated revision). In [9] we correlated fluency (stated by subjective evaluators) with a set of prosodic correlates related with rhythm and F0 contour evolution and in [10] with the distribution of automatic prosodic

10.21437/WOCCI.2017-4

labels. Nevertheless, the corpus used in these experiments consists on the readings of short sentences in contrast with the longer reading activity that is matter of study in this paper.

2.1.2. L1 learners

Bernstein et al have shown that fluency, automatically computed as number of words correct per minute (WCPM), correlates with the scores assigned by human evaluators both for non-native [11] and children's oral reading [12] activities. WCPM is also used by Bolanos et al [13] with the same goal. In this work, we show that speed is not the only aspect to be taken into account for the evaluation of reading fluency.

Concerning the evaluation of children reading skills, Mostow et al [14] present a tutor system for assisting children during reading based on the analysis of the output of an automatic recognition system (ASR) [15]. In [16], the problems raised by the accurate and robust recognition of children's speech is reported. More recently, Proença et al [17] propose another specialized ASR system with the same goal. Both publications are more involved in solving the problem of the recognition of disfluent speech than in correlating results with subjective scores of the quality of the reading.

2.2. Subjective assessment of fluency

When asking teachers on what base they judge a reader as fluent, the answers show a great diversity [18]: accuracy, speed, joy, confidence, expressivity, good comprehension, attention to punctuation, appropriate intonation... Scales have been developed to give useful tools to standardize this subjective assessment.

2.2.1. Accuracy and automaticity

The accuracy can be easily assessed by listening to oral reading and counting the number of error in a list of words [19]. The analysis of error types can help the teacher identify the failing strategies of the pupil. Automaticity is mostly assessed by measuring a reading speed expressed as the number of word read per minute. The more reliable way of assessing the reading rate is using a meaningful text (instead of a word list [20]) and aloud reading [21] in a timed task.

The traditional fluency assessment combines accuracy and automaticity assessments by using the sole reading rate - e.g. National Reading Panel [22], CBM test [23] in English, ELFE [24] in French. The reading rate correspond to the WCPM. It is collected as follows:

Reading The child is instructed to read a standardized text as correctly and fast as possible during one minute.

Assessing The assessor circle every disfluency, i.e. substitution, adding and mispronounciation, and cross the omitted word while the child is reading. He asks the child to stop after one minute and marks the last word read.

Rating The assessor calculates the reading rate by substracting the number of disfluencies from the number of word effectively read.

It is to be noted that this assessment is done online and the number of identified disfluences can differ from one assessor to another.

2.2.2. Prosody

While accuracy and automaticity have long been assessed to evaluate the reading level of pupils, prosody is a more recent adding to the fluency assessment [3]. The prosody can be measured through the oral reading of a meaningful text and rated subjectively by assessors. According to Hudson et al. [19], assessors must observe the appropriate use of prosody markers, such as emphasis, voice tone, inflection, vocal tone in dialogues, phrase boundaries and syntactic pauses (e.g. subject-verb division, preposition, conjunction). In 1991, Zutell and Rasinsky developed quantifiable scales based on these observations and a group of teachers tested them in assessing conditions [18]. They used 2 unidimensional scales of respectively 6 and 4 levels and a multidimensional scale, 3 parameters with 4 levels, based on pace of reading, phrasing and prosodic features (stress, intonation, duration). In 1995, Pinnel et al. [25] proposed a rating scale, currently used by the National Assessment of Educational Progress (NAEP), and presented in table 1. This scale has 4 levels of fluency mostly based on pauses and phrasing.

More recently, Rasinsky suggested 2 other rating scales: a unidimensional scale based on Pinnel's work and a multidimensional based on his previous work with Zutell [18]. Rasinsky add the notions of pace and expressivity to the pause and phrasing assessment of the NAEP scale. The multidimensional scale, presented in table 2 is also based on this 4 parameters : expression, phrasing, smoothness and pace, rated from 1 to 4. The assessment gives a score between 4 and 16. According to Rasinsky, a score below 8 indicate that fluency may be a concern. It is to be noted that this assessment is also performed online and the score may differ from one assessor to another.

3. Data and subjects

3.1. Readings

We collected readings of the French text Alouette" [26] by 116 children recorded either in the classroom (70%) or during language assessment sessions conducted at the Grenoble hospital by speech therapists. This text consists of 265 words. It contains several obstacles (infrequent words, words with high frequency competitors, etc). It is surely not the kind of text to which children are usually confronted with, but the poor predictability of its words emphasizes the child's automaticity level. The WCPM has been calibrated using several hundreds of children readings and delivers an estimation of the reading age. This test is largely used in French-speaking countries to detect delays in reading acquisition and developmental dyslexia. Most children are 8-9 years old and are in second or third grade (see Figure 1). Children were instructed to correctly read aloud the text at a comfortable speech rate. We automatically stopped the recordings after 3 minutes (or before if the child has finished the reading this only occurred 2 times).

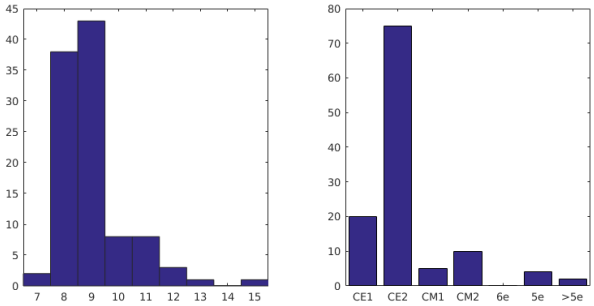

Figure 1: *Histograms of ages and grades of the children. CE1 in France corresponds to the $2^{n.d}$ grade.*

Table 2: *Multidimensional fluency scale (adapted by Rasinski from [18]).*

Dimension/Score	1	2	3	4
Expression and Volume	Reads in a quiet voice as if to get words out. The reading does not sound natural like talking to a friend.	Reads in a quiet voice. The reading sounds natural in part of the text, but the reader does not always sound like they are talking to a friend.	Reads with volume and expression. However, sometimes the reader slips into expressionless reading and does not sound like they are talking to a friend.	Reads with varied volume and expression. The reader sounds like they are talking to a friend with their voice matching the interpretation of the passage.
Phrasing	Reads word-by-word in a monotone voice.	Reads in two or three word phrases, not adhering to punctuation, stress and intonation.	Reads with a mixture of run-ons, mid sentence pauses for breath, and some choppiness. There is reasonable stress and intonation.	Reads with good phrasing; adhering to punctuation, stress and intonation.
Smoothness	Frequently hesitates while reading, sounds out words, and repeats words or phrases. The reader makes multiple attempts to read the same passage.	Reads with extended pauses or hesitations. The reader has many rough spots.	Reads with occasional breaks in rhythm. The reader has difficulty with specific words and/or sentence structures.	Reads smoothly with some breaks, but self-corrects with difficult words and/ or sentence structures.
Pace	Reads slowly and laboriously.	Reads moderately slowly.	Reads fast and slow throughout reading.	Reads at a conversational pace throughout the reading.

Table 3: *Excerpt of the pronunciation dictionary for the word "cassettes". _ and __ respectively stand for syntactic vs respiratory pauses. Note that internal pauses may be also encountered.*

nb	word	pronunciation
...	...	...
19	CASSETTES	k a s e^ t
12	CASSETTES	k a s e^ t __
2	CASSETTES	k a s e^ t q
1	CASSETTES	k a s e^ t x^
1	CASSETTES	k a s e^ t x^ __
1	CASSETTES	k a s __ s e^ t __
...	...	...
4	CASSETTES*	k a s k e^ t
2	CASSETTES*	k a s e^ __
2	CASSETTES*	k a s k e^ t __
1	CASSETTES*	k a __
1	CASSETTES*	k a rX _ s e^ t _
1	CASSETTES*	k a s __
1	CASSETTES*	k a s a __
1	CASSETTES*	k a s k __
1	CASSETTES*	k a s k e^ __
1	CASSETTES*	k a s k e^ t x^ __
1	CASSETTES*	k a z e t
1	CASSETTES*	s^ o s e^ t
...	...	...

3.2. Speech processing

The speech of the children was aligned with a statistical model whose phonetic triphone models, pronunciation dictionary and trigram model were constantly updated using HTK [27] and SLIRM [28] toolkits. While most speech recognizers consider mispronunciations and disfluencies – such as false starts, repetitions, etc. – differently from standard entries into the pronunciation dictionary [16], we treat correct, incorrect or incomplete words the same way in the pronunciation dictionary and the language model. The pronunciation is thus quite large (here an average of 13.4 pronunciation variants per lemma) and the language model implicitly captures syntactic constraints on disfluencies, e.g. false starts and disfluent enunciations often precede the correct one if any.

The labeling of words was thus performed with the following principles: (a) a new word is considered as initiated when at least one vocalic nucleus has been spelled; (b) a star is appended to any incorrect or incomplete word; (c) each dictionary entry begins with a phonated sound; (d) syntactic and respiratory pauses are considered as part of the preceding word. The 100 readings of these 265 words result in 1851 different correct vs. 1697 incorrect/incomplete pronunciations (see excerpt in table 3). Note that 72 correct vs. 123 incorrect/incomplete entries comprise internal pauses.

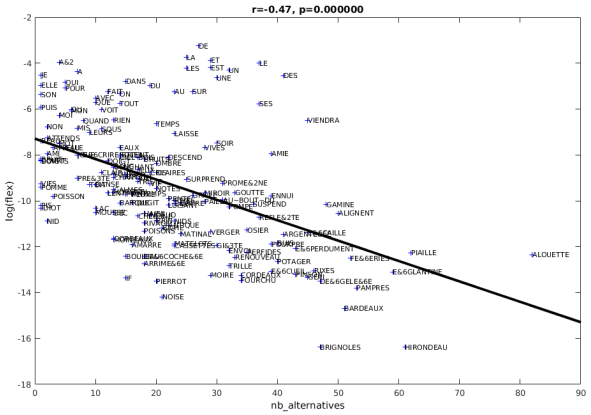

Figure 2: *Log(lexical frequency+0.01) as a function of number of pronunciation alternatives for each lexeme of the corpus. The correlation is highly significant.*

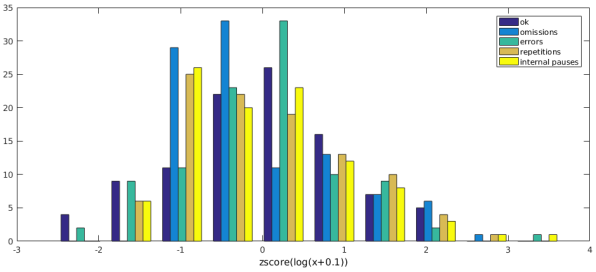

Figure 3: *Distributions of our 5 normalized objective characterizations of the 116 readings of the Alouette. A prior $\log(x+0.1)$ transformation is applied.*

4. Objective characterization

Figure 2 shows that the number of alternative pronunciations (incl. correct vs. incorrect pronunciations, false starts, hesitations, etc.) for each lexeme of the corpus does negatively correlate with the logarithm of the lexical frequency of these words (extracted from [29]): access to the pronunciation of low-frequency words mostly triggers competing high-frequency words or resorts to improper letter-to-sound rules. Lexical frequency is thus expected to strongly impact the occurrence of disfluencies.

We further performed the alignment of the uttered words with the original text and computed the following features (number per minutes): correct words, omitted words, incor-

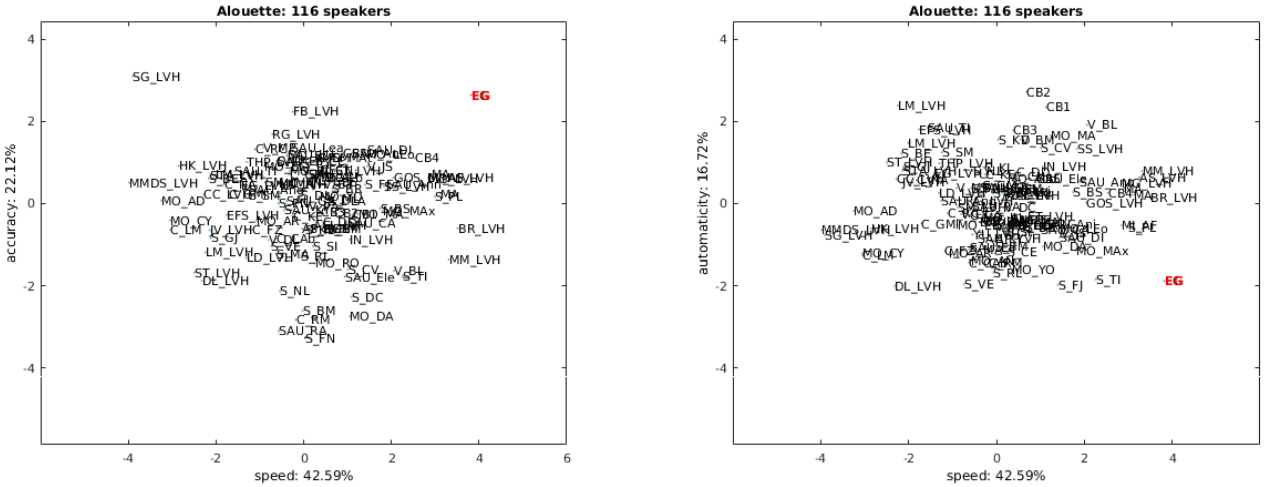

Figure 4: *Projection of our objective characterization of the child readings onto the three main principal axes. Left: 2nd vs. 1st; Right: 3rd vs. 1st. Note that a prior $log(x+0.1)$ transformation is applied to get Gaussian distributions (see Fig.3). The reading of one adult EG is added as a reference and displayed in red. See the text for the interpretation of these axis.*

rect words, repeated words, word-internal pauses. Figure 4 displays the projection of these 5 objective features – with prior log(x+0.1) transform to comply with Gaussian distribution (see Figure 3) – onto the two main principal axes. The first principal axis explains 39% of the data variance and is mainly correlated to the three features: correct words ($r = .82$), word-internal pauses ($r = .81$), incorrect words ($r = .55$). The second principal axis explains 23% of the data variance and is mainly correlated to the rate of omitted words ($r = .84$): children `C_RM` and `CMO_DA` at the top of the figure omitted respectively 26 and 22 words while pronouncing 180 and 218 words correctly. The third principal axis explains 22% of the data variance and is strongly correlated to the rate of repetitions ($r = .89$). A closer examination of the characteristics of the outliers clarifies the interpretation of the main axis of variation:

Speed The first principal axis is effectively related to WCPM. Our adult fluent speaker `EG` (176 WCPM) is effectively located at one extreme position of the first axis while `MMDS` (19 WCPM, 38 pronunciation errors and 19 internal pauses) and `SG_LVH` (28 WCPM, 23 internal pauses) struggle with word spelling.

Accuracy The second principal axis is mainly correlated with accuracy. Again our adult fluent speaker `EG` is located at one extreme position of this second axis. Imprecise speakers are located at the bottom of the map: they combine high rate of errors and omissions (e.g. `S_FN` (68 WCPM) with 14 omissions and 87 errors).

Automaticity The third principal axis is mainly correlated with number of repetitions. Speakers `CB2`, `CB1`, `LM_LVH` repeated some words 32, 29 and 22 times before getting to the next ones.

We here enlighten two different reading strategies adopted by children with reading difficulties: either read carefully and struggle with slow syllabic decoding or rely on lexical, syntactic or semantic bootstrapping to rapidly go ahead at the expense of large number of omissions and errors. We expect both strategies to equally impair comprehension both for the reader and listeners. This issue will be addressed in the near future.

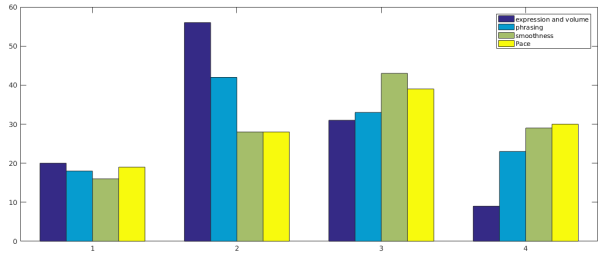

Figure 5: *Distributions of the four subjective ratings into their four levels of abilities.*

5. Subjective assessment

The subjective rating of the 116 recordings was performed using the Rasinsky's multidimensional scale (see Table 2). The rating was done by two assessors familiar with the subjective rating of fluency of 3rd to 5th graders. Only the first minute of each recording was used for the rating. The assessor listen 4 times to the record, one for each ability to rate : expression and volume, phrasing, smoothness and pace. Each ability was given a mark between 1 and 4 according to the scale. Note that pace and smoothness are easy to rate subjectively, whereas expressivity and phrasing are more complex to assess, mostly because of their variations along the text. Figure 5 displays the distributions of the four subjective ratings into their four levels of abilities. We here rate pupils from 8 to 12 years: their expressivity is rather low while pace receives better ratings.

The inter-labeler agreements are average but all significant: linear weighted Cohen's Kappa coefficients are .38, .51, .50 and .42 respectively. In the following, subjective ratings are nevertheless taken as the average of the ratings of two assessors.

We performed a multinomial logistic regression between objective measurements (expressed as `log(nb per mn+0.1)`) and subjective ratings using a Leave-One-Out procedure. The Spearman correlation coefficients (Scc) are respectively .7, .81, .81 and .85 (see Figure 6) while mean absolute errors are close to .5, i.e. respectively .54, .56, .52 and .51. All correlation coefficients are highly significant. We are clearly missing objective features measuring

expressiveness and volume of the voice that cannot be deduced from pronunciation correctness. Adding speech rate (expressed as number of syllables per mn) and F0 variations (expressed in cents) only slightly increases the prediction of expressivity ratings. We suspect that the quality of expressiveness and volume should take into account more detailed linguistic and paralinguistic information.

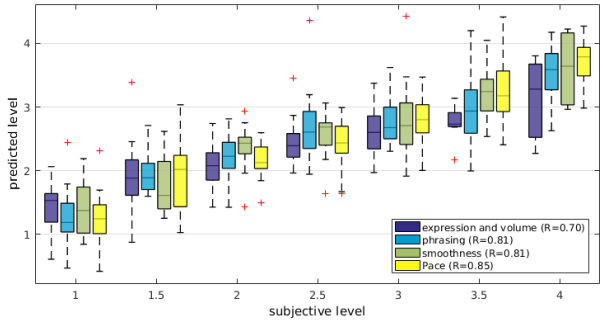

Figure 6: *Boxplots of predicted ratings according to subjective ratings. Legend gives correlation coefficients.*

6. Conclusions and perspectives

We propose here to predict subjective ratings of reading fluency by objective features that can be delivered by an automatic analysis of the verbal content. We are still lacking reliable predictors of reading expressiveness. This is perhaps due to the fact that most of our pupils are still struggling with automaticity and that the target text does not allow for much enthusiasm. We however expect that reliable predictors of reading expressiveness should consider the goodness of fit between the syntactic and semantic content of phrases and their associated prosodic patterns.

7. Acknowledgements

This work is supported by the e-FRAN project Fluence, sponsored by the "Investissements d'avenir" program. We warmly thank Monique Battuz, Catherine Brissaud and Sonia Mandin for providing additional readers.

8. References

[1] T. V. Rasinski, "Assessing reading fluency." *Pacific Resources for Education and Learning (PREL)*, 2004.

[2] T. V. Rasinski, C. L. Blachowicz, and K. Lems, *Fluency instruction: Research-based best practices*. Guilford Press, 2012.

[3] M. R. Kuhn, P. J. Schwanenflugel, and E. B. Meisinger, "Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency," *Reading Research Quarterly*, vol. 45, no. 2, pp. 230–251, 2010.

[4] Z. Breznitz, *Fluency in reading: Synchronization of processes*. Routledge, 2006.

[5] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study." *SLaTE*, vol. 2009, pp. 2–5, 2009.

[6] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *International Symposium on automatic detection on errors in pronunciation training (ISADEPT)*, vol. 6, pp. 1–8, 2012.

[7] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[8] K. Li, X. Wu, and H. Meng, "Intonation classification for L2 English speech using multi-distribution deep neural networks," *Computer Speech & Language*, vol. 43, pp. 18–33, 2017.

[9] V. Cardeñoso-Payo, C. G. Ferreras, and D. E. Mancebo, "Assessment of non-native prosody for Spanish as L2 using quantitative scores and perceptual evaluation." in *LREC*, 2014, pp. 3967–3972.

[10] D. Escudero-Mancebo, C. González-Ferreras, E. Estebas-Vilaplana, and L. Aguilar, "Automatic assessment of non-native prosody by measuring distances on prosodic label sequences." in *Interspeech*, 2017, pp. 1444–1448.

[11] J. Balogh, J. Bernstein, J. Cheng, A. Van Moere, B. Townshend, and M. Suzuki, "Validation of automated scoring of oral reading," *Educational and Psychological Measurement*, vol. 72, no. 3, pp. 435–452, 2012.

[12] R. Downey, D. Rubin, J. Cheng, and J. Bernstein, "Performance of automated scoring for children's oral reading," in *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. ACL, 2011, pp. 46–55.

[13] D. Bolaños, R. A. Cole, W. Ward, E. Borts, and E. Svirsky, "Flora: Fluent oral reading assessment of children's speech," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, p. 16, 2011.

[14] J. Mostow, G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, M. B. Sklar, and B. Tobin, "Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction," *Journal of Educational Computing Research*, vol. 29, no. 1, pp. 61–117, 2003.

[15] S. Banerjee, J. E. Beck, and J. Mostow, "Evaluating the effect of predicting oral reading miscues." in *InterSpeech*, 2003, pp. 3165–3168.

[16] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, 2007.

[17] J. Proença, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigao, "Detection of mispronunciations and disfluencies in children reading aloud," in *Interspeech*, 2017, pp. 1437–1441.

[18] J. Zutell and T. V. Rasinski, "Training teachers to attend to their students oral reading fluency," *Theory Into Practice*, vol. 30, no. 3, pp. 211–217, 1991.

[19] R. F. Hudson, H. B. Lane, and P. C. Pullen, "Reading fluency assessment and instruction: What, why, and how?" *The Reading Teacher*, vol. 58, no. 8, pp. 702–714, 2005.

[20] J. R. Jenkins, L. S. Fuchs, P. Van Den Broek, C. Espin, and S. L. Deno, "Accuracy and fluency in list and context reading of skilled and rd groups: Absolute and relative performance levels," *Learning Disabilities Research & Practice*, vol. 18, no. 4, pp. 237–245, 2003.

[21] L. Fuchs, D. Fuchs, S. Eaton, and C. Hamlet, "Relations between reading fluency and reading comprehension as a function of silent versus oral reading mode," *Unpublished raw data*, 2000.

[22] N. R. P. (US), N. I. of Child Health, and H. D. (US), *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health, 2000.

[23] M. K. Hosp, J. L. Hosp, and K. W. Howell, *The ABCs of CBM: A practical guide to curriculum-based measurement*. Guilford Publications, 2016.

[24] C. Lequette, G. Pouget, and M. Zorman, "Elfe. évaluation de la lecture en fluence," 2008.

[25] G. S. Pinnell *et al.*, *Listening to children read aloud: Data from NAEP's integrated reading performance record (IRPR) at grade 4*. ERIC, 1995.

[26] P. Lefavrais, *Test de l'Alouette*. Paris: Editions du centre de psychologie appliquée, 1967.

[27] P. C. Woodland, C. Leggetter, J. Odell, V. Valtchev, and S. Young, "The 1994 htk large vocabulary speech recognition system," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 73–76.

[28] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit," in *Interspeech*, vol. 2002, Denver, CO, 2002, pp. 901–904.

[29] B. New, C. Pallier, M. Brysbaert, and L. Ferrand, "Lexique 2: A new french lexical database," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 516–524, 2004.