Reading Prosody Development:
Automatic Assessment for a Longitudinal Study

Erika Godde[1,2], Gérard Bailly[1], Marie-Line Bosse[2]

[1] Univ. Grenoble Alpes, CNRS,Grenoble INP, GIPSA-Lab, 38000 Grenoble, France
[2]Univ. Grenoble Alpes, CNRS, LPNC, 38000 Grenoble, France

`[erika.godde; gerard.bailly]@gipsa-lab.fr, marie-line.bosse@univ-grenoble-alpes.fr`

Abstract

We discuss here the benefit of an automatic assessment technique, recently proposed for multidimensional scoring of children readings. The objective of this research is to monitor the development of reading prosody in a longitudinal study. We recorded 57 children in grade 2, then during their grade 3 while reading the same text. First year recordings were assessed subjectively and used to train the automatic tool. The second year recordings were also assessed both subjectively and automatically. We compare here the results given by the automatic vs. subjective assessment and have a look at the progress of the assessed reading skills of the children : expressivity, phrasing, smoothness and rate.

Index Terms: prosody, reading, automatic assessment, longitudinal study, development

1. Introduction

Reading curriculum is often focused on the acquisition of decoding and automaticity. So many fluency training or intervention are based on reading speed and accuracy. However focusing on reading speed is often counterproductive [1]. In fact, when focusing of the number of correct words pronounced per minute, children tend to read too fast, with only few pauses. Breath pauses are often not planned correctly: they breath when they get out of air [2]. The gain in terms of speed is thus performed at the expense of comprehension, creating what Meisinger et al. [1] named "word callers", i.e. good decoders, poor comprehenders. Thus, definitions including prosody as a part of reading fluency skills have been recently proposed [3, 4, 5]. In these definitions, the first stages of reading acquisition, i.e. decoding and automaticity are completed with the desirable acquisition of phrasing and expressivity. According to the four NAEP stages of reading acquisition [6], the first and second stages are decoding and automaticity. Then the third and fourth stages concern prosodic skills. When cognitive resources are freed by decoding automaticity, they can be fully dedicated to prosody and comprehension [7]. So there should be a strong link between the production of adequate prosodic patterns and fluency[8].

Prosody in reading is also related to comprehension and overall literacy in primary and secondary grades, e.g. [9, 10, 11]. Moreover, the link between prosodic reading and reading comprehension strengthens when the reader matures [12]. This put forward the value of interventions on reading prosody in the reading curriculum and literacy achievement [13]. That's why it is important to have a better knowledge of reading prosody development to design accurate and useful intervention.

Figure 1: *Multidimensional Fluency Scale from Rasinski [4].*

2. State of the art

There are few studies on reading prosody development itself, and these studies often considered one or two grade levels only, e.g. [14, 15]. However these studies give an idea of a global development scheme [8]. But to have a better insight of the development of each reading parameter (prosodic or not) and their relations, there's a need of longitudinal studies.

In the literature, longitudinal studies about reading prosody have essentially explored two aspects:

Prosodic patterning Studies have been focused on the patterning of prosodic features through time and space : pauses (number, duration and grammaticality) and variations of syllabic duration, range and slope of pitch variations, match between child and adult F0 contours [16] . . .

Prosody in relation with other skills Longitudinal studies have also explored correlation analysis between prosody and subjective ratings of the mastering of various skills, such as comprehension or fluency, e.g. [17].

In both cases, longitudinal studies enable to test models of development. To our knowledge, there's only one longitudinal study focusing on the development of reading prosody. Miller et al. [16] showed that the rate of pausal intrusions in 1st grade is a good predictor of F0 match in 2nd grade. Moreover 1st and 2nd grade F0 match is related to fluency in 3rd grade. These results suggest that a decrease of pausal intrusion is a precursor of improved intonation and the early development of intonation contour plays an important role in the development of later stage fluent reading. Studies about prosody's relation to other reading skill are more frequent. For example, to our knowledge, there are five studies focusing on the causal relationship between reading prosody and comprehension [8]. Those studies aren't all conclusive. The possible reasons of this inconclusive results are methodological issues or the discrepancy between readers' mother language, age or reading level. Nevertheless the synthesis of these studies suggests that the relation between

reading prosody and comprehension is changing over time and the maturation of reading skills. Thus longitudinal studies enables to trace relation over time and gives us a better understanding of the predictor of possible future reading difficulties. It then could point out the best time to train particular skills in order to avoid future reading difficulties.

These longitudinal studies imply different types of assessment. When the interests are the prosodic features themselves, like pausal intrusion or intonation contour match to expert readers, the author used acoustic analysis of their reading aloud recordings [16]. Several acoustic parameters have been identified as characteristic of prosodic marking [18] although this kind of analysis is time consuming. Automation of acoustic analysis is a subject of interest [19]. When the interest is the evolution of relations, authors prefer to use subjective assessment scales because of their ease and speed of implementation. Figure 1 presents the Multidimensional Fluency Scale, widely used to assess reading prosody in the literature [4]. Subjective scales give access to a global prosodic score or dimension-specific ratings (e.g. expressivity, phrasing...). Note that the subjective assessment is closely linked to the patterning of acoustic parameters [18]. However, in the case of subjective assessment, the inter-rater agreement is highly dependent on the number of raters and their joint training [20]. Automatic assessment aims at cutting out these limitations. It is currently our object of investigation [21, 22].

We propose here to study the evolution of prosodic skills in the first years of reading acquisition, from 2nd to 3rd grade, using an automatic assessment of reading fluency, exploiting both linguistic and prosodic features [22]. We hypothesized that the scores on the different prosodic parameters will increase from 2nd (G2) to 3rd (G3) grade. We also hypothesized that the increase will depend on the fluency scores of the children: good readers and poor readers will make less progress than average readers.

3. Method

3.1. Participants

As part of a larger study, 67 children were recorded in G2 (June 2018) (age: 7 years 11±4 months) and 62 were recorded again in G3 (February 2019) (age: 8 years 7±4 months). The children were recorded at their schools (2 primary schools from Grenoble area) with the authorization of the schools directors and their parents. Considering that some pupils left the school between G2 and G3 grade and some others arrived, only 57 children were recorded both in G2 and G3 grade: 30 boys and 27 girls.

The children were representative of their grade levels both in terms of fluency - in G2 and G3, according to the fluency test Evaleo 6-15 [23] (G2:119.3±41 words, G3: 175.7±53 words) - and non-verbal IQ in G2 - according to the RAVEN's matrix test of non-verbal IQ (G2:33.75 ±9.4) [24].

3.2. Experimental Settings

The recordings were performed individually in a separate quiet room. The children were asked to read a 174-words narrative text written by the authors based on a French short story for kids. There was no particular difficulty in terms of lexicon and syntax for primary grade children. The subjects were asked to read as they were reading a story to a preschooler. The same text was proposed in both years. Considering the delay between the two recordings and the age of the children, there is no test-retest effect. We recorded their aloud voice using a Schur Beta
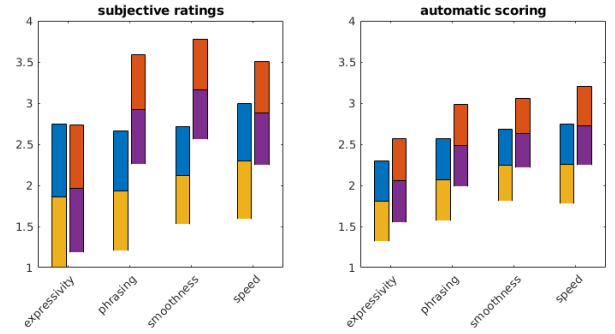

Figure 2: *Comparison of average and standard deviations for each rated dimension and for each year. Left: subjective ratings; right: predicted scores.*

53 microphone and a Berhinger MIC100 amplifier.

grade	coeff	EXP	PHR	SMT.	PAC	total
G2	IRA	.72	.56	.49	.75	.87
G2	corr	.72	.81	.65	.84	.85
G3	IRA	.50	.75	.42	.46	.71
G3	corr	.63	.77	.78	.70	.79

Table 1: *Interrater agreement (IRA) for each subjective dimension assessed, calculated with Krippendorf's alpha for ordinal data [25] and correlation coefficient between subjective and automatically predicted scores.*

4. Multidimensional Fluency Assessment

The recordings of G2 and G3 were assessed subjectively by three trained raters. Data from 237 oral readings (3100 utterances, 5h23') of children of various levels (9 G2, 54 G3, 50 G4, 58 G5, 30 G6 and 36 G7) – not comprising those studied here – are used to train an automatic scoring system. This system was then further used to predict progress.

4.1. Subjective assessment

3 assessors assessed the 114 readings using the multidimensional fluency scale proposed by Rasinski [4] adapted for French (see Figure 1). Each assessor gives a score between 1 (beginner) to 4 (master) on 4 parameters : expressivity (EXP), phrasing (PHR), smoothness (SMT) and pace (PAC). A total score between 4 and 16 is then calculated. The assessors first get trained on a subjective scale with 10 recordings. They discuss their assessments on these recordings to agree on a final score. It enables them to have a clear and unified definition of each parameter assessed in the scale. Then they rate the assessment independently. It is to be noted that an assessor changed between first and second year and that in year 2 one of the rater didn't rate all the recording immediately after the training. Inter-rater agreements calculated with Krippendorf's alpha for ordinal data [25] are presented for each parameter and years in table 1.

Please note that the inter-rater agreement is good for the first year, but lowers for the second year. It can be due to the change in the raters or the interruption in one rater's assessment. It could also come from the decrease of data variability in G3: raters tend to slightly modify their rating criteria to exploit the whole scale. The final subjective scores are the averages of the 3 raters scores. Figure 2 presents the average and standard de-
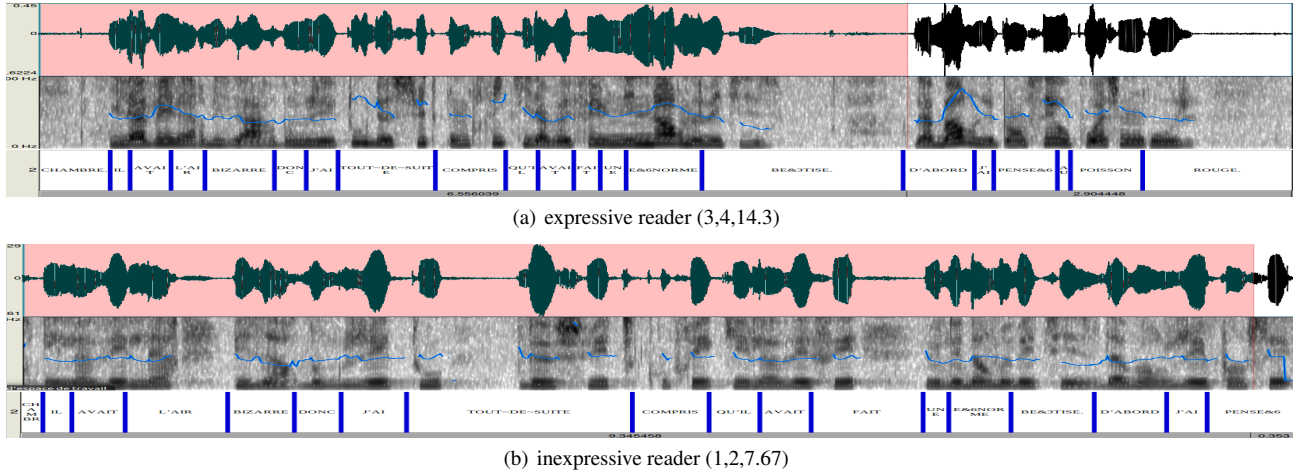
105

(a) expressive reader (3,4,14.3)

(b) inexpressive reader (1,2,7.67)

Figure 3: *Example of expressive (a) and inexpressive (b) readings spectra and pitch variations visualized with Praat. Both readers read the same sentence, scores are indicated as follows (EXP, PHR, total). Reader a breathes at full stops (after "chambre.", "btise", "rouge.") while reader b doesn't and inhales at illicit places (after "l'air" and "fait" and inside the adverbial phrase "toutde suite").*

viations for each rated dimension and for each year.

Figure 3 illustrates the relation between acoustic parameters and subjective ratings with two recordings, visualized with Praat. Figure 3(a) shows the recording of an expressive reader, rated 3 out of 4 in EXP and 4 in PHR, for a total of 14.33 out of 16. It shows a wide amplitude of pitch variation and few grammatical pauses. On the contrary, the inexpressive reader in Figure 3(b), rated 1 in EXP, 2 in PHR for a total of 7.67, performed a reduced pitch variation and produced ungrammatical pauses while ignoring punctuations.

4.2. Automatic assessment

Using data from 157 children recorded either in the classroom (70%) or during language assessment sessions conducted at the Grenoble hospital by speech therapists, we demonstrated [21] that subjective ratings could be predicted from speech signals with a rather good accuracy. In fact, we proposed new features to enhance the prediction of expressivity [22] that exploits a projection of the child's F0 and rhythmical contours onto a prosodic manifold computed from multiple (i.e. 20) adults readings. In fact, we used multidimensional scaling (MDS) to first represent prosodic patterns of the adult stimuli as points in a n-dimensional space. Prosodic patterns of children reading will be then represented by their projection onto this reference space. The loading factors of the projection of F0 and rhythmical contours will be called respectively CF_i and CR_i in the following. We only consider projections onto the three first MDS planes (i.e. $i = 1, 3$).

As part of a larger study, we also recorded 237 pupils from G2 to G7 grades reading the same text used in the present study. For each childs reading, we performed the text-to-speech alignment using an extended lexicon with correct and incorrect pronunciations together with a disfluent bi-gram model (i.e. incorporating false starts, repetitions, and incorrect pronunciations). These disfluent lexicon and bi-gram model are trained on the children' readings [21]. We then computed the following features (number per minutes): correct words (CWPM), omitted, incorrect and repeated words (IWPM), number of vowels (VPM). We then performed a multinomial logistic regression of ordinal data (ordinal R package) between these objective measurements (expressed as log(nb per mn+0.1)) and subjective rat-

ings. Although CWPM, IWPM and VPM are good predictors of PHR, SMT and PAC, they are insufficient to accurately estimate EXP. When adding prosodic features (CF_i and CR_i), the predictions of EXP and PHR significantly improve and reach the same level of performance as for SMT and PAC.

We further performed model simplification by iteratively removing predictors whose χ^2 value is less than 0.05. We end up with the following formula (predictors are given with decreasing significance):

PAC~1+CWPM+CR_1
SMT~1+VPM+IWPM+CR_1
PHR~1+CR_1+VPM+CF_3+CR_3+CR_2
EXP~1+CF_1+CR_1+VPM+CR_2

In the following, we differentiate human scores vs. the predictions from objective measures by adding a wide hat on the predicted scores.

4.3. Comparison between subjective and automatic assessment in G3

The right part of Figure 2 presents the average and standard deviations of predictions for each rated dimension and for each year. Compared with subjective scores figured on the left part, predictions clearly underestimate the progress accomplished by children, in particular on PHR and SMT. While the progress between grades is clear and significant in both objective and subjective terms, human raters give a bonus of 0.5 points to both scales. This is particularly surprising for PHR since its IRA is rather high (i.e. 0.75 in Table 1). We could think of several hypothesis regarding this gap:

- an effect of the discrete scoring framework. Improvement is rated with one point increase in the subjective scale, while it is more finely observed by the automated continuous rating.
- an effect of the lower inter-rater agreement. One rater's scoring tends to be higher than the others, increasing the mean scores.
- the human raters only listen to the first minute of oral reading, while the automated scoring is realized on the whole text. Indeed less fluent readers' prosody is better at the beginning and tend to decrease along the text [26].

This discrepancy between data and model is for now unsolved and deserves further investigation.
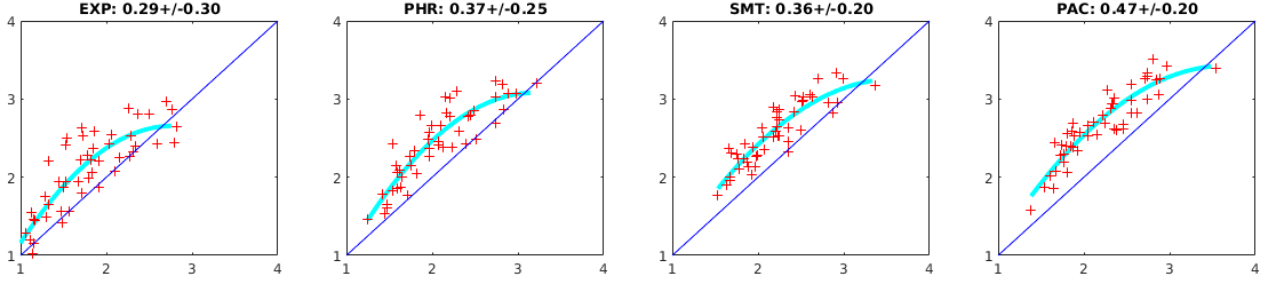
Figure 4: *Estimated G3 scores as a function of G2 scores for each subjective dimension. A second-order polynomial fit is superimposed.*
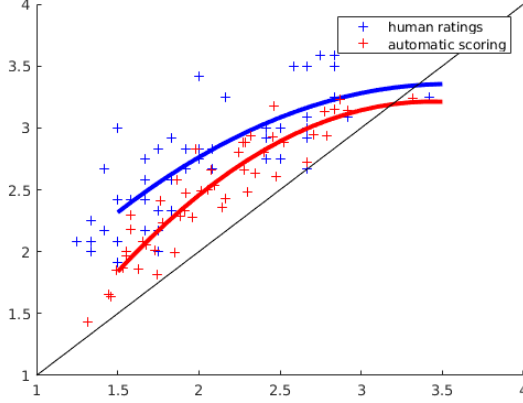

Figure 5: *Mean G3 scores as a function of mean G2 scores. Automatic scoring underestimate subjective progress but tendencies are similar.*

5. Results

Figure 4 shows the predicted scores of grade 3 according to the predicted scores of grade 2. Please note that (1) pupils make progress: there is a general increase in each score. The largest average increase concerns $\widehat{PAC}$, the smallest being $\widehat{EXP}$; (2) medium readers achieve the largest increase of performance (see Figure 5). Poorer readers and skilled readers do not progress so much. Note of course that progress is scale-dependent. Considering mastering automaticity (PAC and SMT) as a perquisite for increasing PHR and EXP, poorer readers undoubtedly make significant progress on PAC while remaining poor on EXP.

In order to examine the conditioning between the performance of pupils at grade 3 and their scores at grade 2, we performed a regression between paired scores at both grades to predict predicted grade 3 scores from grade 2 ones. We also performed model simplification by iteratively removing predictors whose χ^2 value is less than 0.05. We end up with the following formula (predictors are given with decreasing significance):

$$\widehat{EXP_3} \sim 1+\widehat{EXP_2}+\widehat{SMT_2}$$
$$\widehat{PHR_3} \sim 1+\widehat{EXP_2}+\widehat{PAC_2}$$
$$\widehat{SMT_3} \sim 1+\widehat{EXP_2}+\widehat{SMT_2}$$
$$\widehat{PAC_3} \sim 1+\widehat{EXP_2}+\widehat{SMT_2}$$

Please note that the grade 2 expressivity ($\widehat{EXP_2}$) is the first best predictor for all scores at grade 3, while the last symptom of the acquisition of automaticity ($\widehat{SMT_2}$) triggers significant improvements in all scores at grade 3 except phrasing $\widehat{PHR_3}$.

Because of the poor prediction of subjective PHR_3 and SMT_3 ratings by objective features extracted from reading performances, we performed the same operation with subjective scores using the `clmm` function of the R package `ordinal` including raters as a random effect. We obtain the following formula (predictors are given with decreasing significance):

$$EXP_3 \sim 1+EXP_2+PAC_2$$
$$PHR_3 \sim 1+EXP_2$$
$$SMT_3 \sim 1+EXP_2+PAC_2$$
$$PAC_3 \sim 1+EXP_2$$

Again, the grade 2 expressivity (EXP_2) is the most significant predictor for all scores at grade 3!

These results are in line with Schwanenflugel et al [5] who suggest that prosody may serve to mediate between decoding skills and comprehension to enhance comprehension (p. 122). Frazier et al [27, 28] in fact propose that prosody provides a scaffold for reading comprehension by allowing speech sequences to be held in working memory and subsequently analyzed for further processing: prosody helps children to chunk text into meaningful units, helping readers to decode words, parse and understand sentences.

6. Conclusions

We here sketch a method for relating oral reading performance with the objective characterization of the uttered speech signal. Multidimensional scaling (MDS) is used to characterize the quality of prosodic patterns produced by children given the large variety of licit communicative strategies. We show that these new objective cues significantly contribute to compact and accurate predictive models.

We still need to improve our predictors, in particular when considering an increase of the average level of performance of the children. We in fact suspect that human raters take this a priori information when performing the scoring.

This framework opens the pathway to the automatic assessment of reading fluency and faithful characterization of the development of reading prosody across large time periods.

A longitudinal study is currently conducted that involve an annual screening of cohorts of pupils, in particular to assess the benefit of computer-assisted training of reading using Karaoke [29, 30].

7. Acknowledgements

8. References

[1] E. B. Meisinger, B. A. Bradley, P. J. Schwanenflugel, M. R. Kuhn, and R. D. Morris, "Myth and reality of the word caller: The relation between teacher nominations and prevalence among elementary school children." *School Psychology Quarterly*, vol. 24, no. 3, p. 147, 2009.

[2] F. Grosjean and M. Collins, "Breathing, pausing and reading," *Phonetica*, vol. 36, no. 2, pp. 98–114, 1979.

[3] S. L. Dowhower, "Speaking of prosody: Fluency's unattended bedfellow," *Theory into practice*, vol. 30, no. 3, pp. 165–175, 1991.

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

[5] P. J. Schwanenflugel, A. M. Hamilton, M. R. Kuhn, J. M. Wisenbaker, and S. A. Stahl, "Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers." *Journal of educational psychology*, vol. 96, no. 1, p. 119, 2004.

[6] M. C. Danne, J. R. Campbell, W. S. Grigg, M. J. Goodman, and A. Oranje, "Fourth-grade students reading aloud: Naep 2002 special study of oral reading. the nation's report card. nces 2006-469." *National Center for Education Statistics*, 2005.

[7] P. Suárez-Coalla, M. Álvarez-Cañizo, C. Martínez, N. García, and F. Cuetos, "Reading prosody in spanish dyslexics," *Annals of dyslexia*, vol. 66, no. 3, pp. 275–300, 2016.

[8] M.-L. Godde, Erika Bosse and G. Bailly, "A review of reading prosody acquisition and development," *Reading and Writing*, submitted.

[9] D. D. Paige, T. Rasinski, T. Magpuri-Lavell, and G. S. Smith, "Interpreting the relationships among prosody, automaticity, accuracy, and silent reading comprehension in secondary students," *Journal of Literacy Research*, vol. 46, no. 2, pp. 123–156, 2014.

[10] N. Calet, N. Gutirrez-Palma, and S. Defior, "A cross-sectional study of fluency and reading comprehension in Spanish primary school children: Fluency and Comprehension: Cross-sectional Study," *Journal of Research in Reading*, vol. 38, pp. 272–285, Aug. 2015.

[11] N. J. Veenendaal, M. A. Groen, and L. Verhoeven, "The role of speech prosody and text reading prosody in children's reading comprehension," *British Journal of Educational Psychology*, vol. 84, pp. 521–536, Dec. 2014.

[12] K. Yildirim, T. Rasinski, and D. Kaya, "Fluency and comprehension of narrative texts in turkish students in grades 4 through 8," *Education 3-13*, pp. 1–10, 2018.

[13] T. Rasinski, "Reading fluency instruction: Moving beyond accuracy, automaticity, and prosody," *The Reading Teacher*, vol. 59, no. 7, pp. 704–706, 2006.

[14] M. Álvarez-Cañizo, P. Suárez-Coalla, and F. Cuetos, "Reading prosody development in spanish children," *Reading and Writing*, pp. 1–18, 2017.

[15] P. J. Schwanenflugel, A. M. Hamilton, M. R. Kuhn, J. M. Wisenbaker, and S. A. Stahl, "Becoming a Fluent Reader: Reading Skill and Prosodic Features in the Oral Reading of Young Readers." *Journal of Educational Psychology*, vol. 96, no. 1, pp. 119–129, 2004.

[16] J. Miller and P. J. Schwanenflugel, "A Longitudinal Study of the Development of Reading Prosody as a Dimension of Oral Reading Fluency in Early Elementary School Children," *Reading Research Quarterly*, vol. 43, pp. 336–354, Oct. 2008.

[17] N. J. Veenendaal, M. A. Groen, and L. Verhoeven, "Bidirectional Relations Between Text Reading Prosody and Reading Comprehension in the Upper Primary School Grades: A Longitudinal Perspective," *Scientific Studies of Reading*, vol. 20, pp. 189–202, May 2016.

[18] R. Cowie, E. Douglas-Cowie, and A. Wichmann, "Prosodic Characteristics of Skilled Reading: Fluency and Expressiveness in 810-year-old Readers," *Language and Speech*, vol. 45, no. 1, pp. 47–82, 2002.

[19] D. Bolaños, R. A. Cole, W. H. Ward, G. A. Tindal, J. Hasbrouck, and P. J. Schwanenflugel, "Human and automated assessment of oral reading fluency." *Journal of educational psychology*, vol. 105, no. 4, p. 1142, 2013.

[20] T. Haskins and V. Aleccia, "Toward a reliable measure of prosody: an investigation of rater consistency," *Int J Educ Soc Sci*, vol. 1, no. 5, pp. 102–12, 2014.

[21] E. Godde, G. Bailly, D. Escudero, M.-L. Bosse, and E. Gillet-Perret, "Evaluation of reading performance of primary school children: Objective measurements vs. subjective ratings," in *International Workshop on Child Computer Interaction (WOCCI)*, 2017, pp. 23–27. [Online]. Available: http://dx.doi.org/10.21437/WOCCI.2017-4

[22] G. E. . M.-L. B. Bailly, G., "Predicting expressivity ratings of children readings from the speech signals," in *Interspeech*, Graz, Austria, 2019, p. submitted.

[23] L. Launay, "Evaleo 6-15 : batterie d'valuation du langage oral et crit chez les sujets de 6 15 ans," *Rducation orthophonique*, vol. 55, no. 273, 2018.

[24] J. Raven *et al.*, "Raven progressive matrices," in *Handbook of nonverbal assessment.* Springer, 2003, pp. 223–237.

[25] K. Krippendorff, *Content analysis: An introduction to its methodology.* Sage publications, 2018.

[26] R. G. Benjamin and P. J. Schwanenflugel, "Text complexity and oral reading prosody in young readers," *Reading Research Quarterly*, vol. 45, no. 4, pp. 388–404, 2010.

[27] L. Frazier, K. Carlson, and C. Clifton Jr, "Prosodic phrasing is central to language comprehension," *Trends in cognitive sciences*, vol. 10, no. 6, pp. 244–249, 2006.

[28] C. Clifton, K. Carlson, and L. Frazier, "Tracking the what and why of speakers choices: Prosodic boundaries and the length of constituents," *Psychonomic Bulletin & Review*, vol. 13, no. 5, pp. 854–861, 2006.

[29] G. Bailly and W.-S. Barbour, "Synchronous reading: learning french orthography by audiovisual training," in *Interspeech*, 2011, pp. 1153–1156.

[30] E. Gerbier, G. Bailly, and M. L. Bosse, "Audiovisual synchronization in reading while listening to texts: Effects on visual behavior and verbal learning," *Computer Speech & Language*, vol. 47, pp. 74–92, 2018.