Automatic assessment of oral readings of young pupils

Gérard Bailly, Erika Godde, Anne-Laure Piat-Marchand, Marie-Line Bosse

Speech Communication 138 (2022) 67–79

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic assessment of oral readings of young pupils

Gérard Bailly [a,*], Erika Godde [a,b], Anne-Laure Piat-Marchand [a], Marie-Line Bosse [b]

[a] Univ. Grenoble Alpes, GIPSA-Lab, 38000 Grenoble, France
[b] Univ. Grenoble Alpes, LPNC, 38000 Grenoble, France

ARTICLE INFO

Keywords:
Fluency
Children
Oral readings
Automatic assessment
NAEP
Speech recognition
Prosody

ABSTRACT

We propose a computational framework for estimating multidimensional subjective ratings of the reading performance of young readers from speech-based objective measures. We combine linguistic features (number of correct words, repetitions, deletions, insertions uttered per minute, etc.) with prosodic features. Expressivity is particularly difficult to predict since there is no unique gold standard. We propose a novel framework for performing such an estimation that exploits multiple references performed by adults and we demonstrate its effectiveness using recordings from a large data set of 1063 oral readings from 442 children (more than 30 h of speech), 84 oral readings from 42 adults and 6853 subjective scores delivered by 29 different human raters. We show that robust and accurate estimations of reading fluency can be achieved using combined features. This automatic assessment tool provides teachers and speech therapists with reliable estimates of the maturation of several reading skills.

1. Introduction

An expert reader is often described as someone who reads as he/she talks, that is not only decoding the text accurately, smoothly and at a conversational rate, but also using melody, rhythm, and intensity variations to capture the listener's attention and ease his/her comprehension of the text. So being a good reader refers not only to accuracy and automaticity (even if these skills are an important prerequisite) but mostly to the ability to use prosodic features to render the meaning of the text read aloud. This observation puts into question the reading curriculum: the traditional way to assess reading skills is to measure the reading rate, with the number of words correctly read in a minute. Measuring the reading rate gives a good indication of accuracy and automaticity skills. However, it gives no insight into on phrasing, expressivity, or comprehension. Worse, this way of assessing leads the children to believe that a good reader is a fast reader, and not a reader who understands what he/she is reading. Many young children then read very fast at the expense of the listener's comprehension, and also their own comprehension (Applegate et al., 2009; Meisinger et al., 2009). Consequently, the educational community has shown a growing interest in reading prosody as part of reading fluency and an indicator of reading comprehension. The definition of reading fluency itself has been revised to go beyond the reading rate. Authors like Rasinski et al. (2019) or Kuhn et al. (2010) proposed new definitions of reading fluency that include accuracy, automaticity, phrasing, and expressivity. These four main skills are also parts of the NAEP's Oral Reading Fluency Scale (Pinnell et al., 1995).

Reading prosody has therefore become an important issue in reading education. The assessment of reading fluency, including prosody, has to go beyond the traditional reading rate. The need for reliable assessment tools for reading fluency including prosody is growing, both for research and educational purposes. However, the question is not trivial. Assessing prosody is very subjective. There is no gold standard, but many possible ways of reading a text expressively. Moreover, expressivity assessment is not only speaker-dependent but also listener-dependent: a reader may sound very expressive to one listener but not to another. The tools needed should be reliable to compare young readers to each other, or to regularly assess reading skills in the classroom and identify reading difficulties. The assessment should also be reliable over time to track progress through multiple assessments of the same reader, for example for longitudinal studies or training evaluations.

We propose and evaluate an automatic system that infers multidimensional subjective scores given verbal and prosodic features computed from the speech signal. Verbal features are estimated using speech recognition techniques with a dictionary comprising possible correct and incorrect pronunciations, and a language model that takes into account mispronunciations, repetitions, and insertions. Prosodic features are computed using an original projection technique using multiple adult readings.

* Corresponding author.
E-mail address: gerard.bailly@gipsa-lab.fr (G. Bailly).

https://doi.org/10.1016/j.specom.2022.01.008
Received 15 February 2021; Received in revised form 25 January 2022; Accepted 26 January 2022
Available online 13 February 2022

2. State of the art

While several authors have explored the acoustic markers of reading prosody and proposed subjective scales that incorporate dimensions related to reading prosody, few works have used prosodic features to predict the subjective quality of readings of whole texts.

2.1. Acoustic markers of reading prosody

Acoustic markers of reading prosody have been explored for numerous purposes: distinguishing read versus spontaneous speech (Howell and Kadi-Hanifi, 1991; Blaauw, 1994), finding the most relevant prosodic features for directly predicting performance on other literacy abilities such as comprehension (see Wolters et al., 2020) or subjective scores (see below).

Dowhower (1991), in an early definition of reading prosody, associated mature reading prosody to six main features: appropriate pause intrusion, phrase segmentation and length, phrase final-lengthening, terminal intonation contours, and stress. These acoustic features are mostly driven by rhythmic abilities (e.g., pause, segmentation, lengthening). Expressivity markers, such as pitch and intensity variation, are missing. More recent studies have explored the best acoustic markers to describe reading prosody, both for phrasing and expressivity. Cowie et al. (2002) explored 40 different acoustic markers in recordings of children reading (8–10 years old). The aim was to identify the features with the highest correlation to the subjective perception of prosody: phrasing and expressivity, and also to fluency in general. So they related the 40 acoustic markers measured to the subjective assessment of fluency and expressiveness of the readings. Unsurprisingly, fluency is mostly linked to rhythmic features: pause duration, pause frequency, syllabic rate, and pitch movement frequency. Expressiveness is mainly linked to pitch variation: pitch movement magnitude and duration and their variation from one sentence to the next. All the acoustic markers were cross-correlated, and rhythmic markers were also secondarily linked to expressiveness and the melodic markers to fluency. Thus, there is no real one-to-one correspondence between objective acoustic markers and perceptual dimensions of reading prosody. Benjamin et al. (2013) examined the correlation between two dimensions (*Expressive Intonation* and *Natural Pausing*) of the Comprehensive Oral Reading Fluency Scale (CORFS) with some acoustic markers of the prosody of young readers (7–8 years). The most relevant features were pitch variation for expressivity and pause variables for phrasing. But, once again, there were secondary cross-links between the features. These studies highlight the fact that acoustic markers contribute to several perceptual dimensions and that their relative contributions depend on the skill to be assessed.

Acoustic markers of reading prosody are also used in studies describing reading prosody development (e.g., Benjamin and Schwanenflugel, 2010; Álvarez-Cañizo et al., 2018; Miller and Schwanenflugel, 2008). The acoustic features usually used to assess reading prosody are pause frequency and duration to characterize phrasing, and pitch variation using adults as references, to characterize expressivity (Godde et al., 2020b). Like in previous studies, the use and choice of acoustic markers by researchers reveal the difficulty in assessing expressivity, even objectively. The use of adult readings as references highlights the subjectivity and variability of expressive reading.

Finally, acoustic markers may provide objective and precise measures of reading performance, in particular across multiple evaluations. However, measurements do not enable us to completely get rid of the issue of the variability of expressivity and its perceptual correlates. Moreover, the use of acoustic markers needs a precise alignment between children's recordings and the text being read aloud, together with the annotation of grammatical versus ungrammatical pauses. Young children's readings are full of repetitions, hesitations, errors, and omissions. Therefore the automatic alignment between text and voices must be completed by strict control and many corrections, especially for the youngest readers. It is very time-consuming and limits its use to small populations.

2.2. Subjective scales

Another common method to assess reading fluency including prosody is the use of subjective scales. In this method, one or several raters listen to a child reading and give scores on one or several dimensions. Several types of scales have been proposed in the past decades. Some were one-dimensional, such as the NAEP's scale (Pinnell et al., 1995) or Zutell and Rasinski (1991). These scales had four levels, mainly focusing on grouping skills, from word-to-word reading to syntactically correct phrases. Expressivity is an expert skill that corresponded to the highest level. Mostly designed for teachers' use, these scales enable them to position children in a global developmental scheme. They supposed that reading fluency develops in successive orderly stages: first decoding, then automaticity, then grouping words in meaningful units, and finally expressivity. However reading prosody development is more complex because all these skills develop in parallel and asynchronously (Godde et al., 2020b) and are mildly cross-correlated, as mentioned in 1.1. So it could be difficult to assess all the fluency skills on a one-dimensional scale.

Consequently, multidimensional scales have also been proposed, for example, the Multidimensional Fluency Scale (MDFS, see Fig. 1) (Rasinski, 2004) or the CORFS (Benjamin et al., 2013) already mentioned. In the MDFS scale, the raters independently evaluate 4 items: rate, smoothness, phrasing, and expressivity, each on a scale from 1 to 4, leading to a cumulative fluency score between 4 and 16. The acoustically grounded scale used by CORFS combines the measure of the reading rate and a subjective scale with 2 items: natural pausing and expressive intonation. The multidimensional scales are more precise and enable the assessment of children with different developing schemes or strategies (e.g., an expressive reader with a low reading rate vs. a reader with a high reading rate but who sounds flat and inserts ungrammatical pauses).

These subjective scales provide teachers with a quick method for assessing children in classrooms. They are also commonly used in research to explore the relationships between reading prosody or fluency and other reading skills such as comprehension or reading rate. They are also used in longitudinal studies exploring causal relationships between repetitive assessments (Miller and Schwanenflugel, 2008; Álvarez-Cañizo et al., 2020) or in the evaluation of fluency interventions (Ardoin et al., 2013). These scales are particularly adapted for probing large populations and conducting modeling studies because they are easy to use and give a limited number of scores for prosodic skills. However multiple assessments with the same or different raters raise the question of inter- and intra-rater reliability. To be reliable, these kinds of scales need at least 2 readings per child and 2 judges (Haskins and Aleccia, 2014; Moser et al., 2014). The judges need to follow precise training so they all give scores according to the same mental model (Haskins and Aleccia, 2014; Smith and Paige, 2019). In this case, the inter-rater agreement and the generalizability coefficient are above 0.9 (Smith and Paige, 2019). But the raters' training is long and requires a lot of recordings: 2 times 3 h and 50 recordings for Smith and Paige (2019). In the case of multiple assessments of the same children throughout one or several years, the training and possible change among the raters could be an issue and inter-rater agreement rapidly falls. In that respect, a rating process calibrated by a set of reference raters could ensure the measurement's reliability along with multiple assessments.

2.3. Automatic assessment

While several authors have proposed to evaluate pronunciation accuracy of isolated words or sentences (see Proença et al., 2017, for a rather extensive review), few references can be found that study the contribution of prosody to reading fluency.

Schwanenflugel et al. (2004) tested several models linking decoding speed, reading prosody, and comprehension skills. They recorded 123

	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.

Fig. 1. Multidimensional Fluency Scale (Rasinski, 2004).

2nd- and 3rd-grade children (N = 123) and 24 adults reading aloud a passage containing 7 sentences. They characterized reading prosody using four factors: intersentential pause structure and pause length, sentence-final F0 declination, and child–adult F0 match. Interestingly, the latter feature is a word-by-word correlation between the F0 contour of each child and an *idealized* adult contour. In their footnotes, Schwanenflugel et al. (2004) argued that, except for a few outliers, adults *adhere* to a *target* adult prosody.

Duong et al. (2011) compared two prosodic models to predict the MDFS scores of 10 children who followed a computer-assisted reading program as well as the post-test fluency, comprehension scores and gains of 55 children: (1) a text-dependent template model that consists of the prosodic patterns of one adult reading each target sentence; the patterns of intensity, duration, and melodic variations are computed at the word-level; (2) a generalized model using a text-to-speech synthesizer that provides means and standard deviations for each element of the contour. Generalized models are found to be superior to template models in all prediction tasks, demonstrating the benefit of using a likelihood measure taking into account a distribution of possible realizations rather than a distance to a unique gold standard.

Bolanos et al. (2013) automatically calculated the number of words correctly read in a minute (NCW) and added several prosodic features: speaking rate, sentence reading rate, number of word repetitions, location of the pitch accent, word and syllable durations, and filled and unfilled pauses and their correlation to punctuation marks. Then, they used Support Vector Machines (SVM) to predict the second and fourth level of the NAEP scale (fluent vs. non-fluent) (Pinnell et al., 1995). They used the recordings of 313 1st to 3rd-grade students rated by 2 experts. They reported a machine-to-human Spearman's rank correlation of .86. However, as seen previously, the NAEP scale gives only a global score, not distinguishing between objective (reading rate) and subjective (expressivity) measurements. They claim that current speech technology may provide robust NCW estimates but conclude that "reading fluency scales have not [yet] been grounded in research on reading prosody". It is rather easy to agree on a reading rate whereas the space of acceptable prosodic patterns is much more difficult to define: good reading prosody depends on many factors such as the reader's specific breathing patterns, dialectal variation and interpretation of textual content.

Proença et al. (2017) combine reading speed and numerous verbal features (in particular, rate of various disfluencies) to predict global scores (0–5) of oral readings of isolated sentences and pseudo-words performed by 284 children (6–10 years in L1 Portuguese). Each of 100 teachers rated 15 readings. Some individual features such as CWPM (Correct Words/Characters Per Minute, referred in this paper as NCW) already predict mean ratings with cross-correlations above .9 and an RMSE close to 0.4. Gaussian Process Regression (GPR) using multiple features raises this performance by 10%. This shows that decoding abilities dominate performance when reading isolated sentences and that a global score also focuses the attention of raters to decoding speed.

Sabu and Rao (2018) combined verbal features – computed via a forced alignment of a *canonical transcription* by a speech recognizer – and prosodic features – computed as correlations with ideal contours – to predict subjective ratings of phrasing and prominence. These ratings were performed by three experts. Children's readings (10–14 years in L2 English) of short passages (10 sentences) are rated and predicted with F-scores of .73 and .68 respectively.

In contrast with these works that hypothesize canonical prosodic contours, we use prosodic manifolds reflecting the multiple licit interpretations and thus various correct prosodic realizations of a given passage.

In the following, we first propose to calculate a prosodic space that takes into account multiple references and places the children's reading in this space. There will be then not one gold standard but several references uttered by expert readers. We proposed this scheme in Godde et al. (2017). This automated rating method compared the child's reading to a set of 20 different expert readers to give 4 scores to the items of MDFS. However, the text used in this early experiment ("l'alouette" used by Lefavrais, 1965), widely used for rating automaticity in France, was not an adequate material to assess expressivity. Indeed the text was difficult and almost meaningless for the children, so very few of them were expressive when reading . . . including the expert readers. We thus collected new reading material. We secondly describe how we combine linguistic and prosodic features to predict the MDFS scores provided by human raters.

3. Speech data collection

We here use 1084 readings of 4 short texts (see Appendix) that were collected for different purposes:

Text 01 This short story is adapted from Friot (2007). This is a 159 word narrative text, with 13 sentences ranging from 7 to 18 words. It was used in a 3-year longitudinal study on reading prosody development. This study took place in two primary schools and one middle school in the Grenoble urban area. The protocol was approved by an ethical committee and the local representative of the French Ministry of Education. In year 1, we recorded 330 pupils from grade 2 to grade 7. Whenever possible, the 61 s graders were recorded again in grades 3 and 4 during the next two years. Only 50 of them recorded the text three times.

Text 02 This text is a 155 word dialog between 2 characters. It contains 23 sentences from 2 to 17 words, 10 of them are questions and 4 are exclamations. The text has no lexical difficulties that could induce decoding issues so that the children can concentrate on phrasing and expressivity and not on decoding. This short dialog was recorded by the same 316 pupils that recorded text 01. The aim was to favor children's expressivity

Text A This short story of 271 words was used in a semi-supervised computer-assisted reading program as a baseline of pre- versus post-training performance. Readers are children from grades 3 to 5.

Text RoiGourmand This longest text of 421 words from Porte and Capek (1998) has been used for reading assessment (Zorman et al., 2008). It was here used as the first exercise in an autonomous computer-assisted reading experiment performed during 10 weeks in first-grade classrooms. We only selected items with Signal-to-Noise Ratio (SNR) superior to 20 dB, estimated using the DeepXi end-to-end solution proposed by Nicolson and Paliwal (2019).

3.1. Adult references

We also recorded the oral readings of 42 adults. 21 adults recorded readings of texts 01 and 02. Another set of 21 adults recorded readings of texts A and RoiGourmand. These readings will be used to build and mark out the prosodic spaces described in Section 4.3.1.

3.2. Protocol

Texts 01, 02, and A were presented on an A4 sheet written in Arial font, size 14, and double interlining. The children were recorded one at a time in their schools, in a quiet room near their classrooms, and under the supervision of the first experimenter. They were instructed to read "as if they were reading a story to a preschooler". During reading, we recorded the voices using a Schur Beta 53 microphone with a Behringer MIC100 amplifier. We also recorded their respiratory movement via abdominal and thoracic belts to study the impact of reading skills on breathing patterns (see analysis in Godde et al., 2022).

The text RoiGourmand was displayed on a 14″ tablet with single spacing. It was recorded during group activities of pupils in the classroom under the supervision of their teacher. Depending on numerous factors (number of children, time of day . . .), the SNR often drops below 0 dB. Despite the use of gaming headsets with close microphones, few clean recordings are available.

Adults were recorded in an experiment room in the lab. They also received the same reading instructions mentioned above.

3.3. Human subjective scoring

We adapted the MDFS (see Fig. 1) for French readers (Godde et al., 2020a) to subjectively assess the children's recordings. Raters listen to the first minute of each recording and then have to rate the 4 items independently: pace (PAC), smoothness (SMT), phrasing (PHR), and expressivity (EXP). Each item receives a score between 1 (no skills), and 4 (expert skills). The rating session involved several raters. It begins with a reading of the scale used and some explanation to make sure that everyone agrees on the definition of each item. As a training, the raters listen to 10 to 15 readings from different levels and deliberate about their ratings. Then they assess all the other readings independently.

In the following, we define bad vs. good readers according to their average score over the 4 dimensions, respectively < 1.5 vs. > 3.5.

We gathered 6853 ratings assigned by 29 raters. Most ratings (62.7%) were given by 6 experts in assessing children's reading (experienced teachers and speech therapists). We also asked non-experts (11 Master students in Psychology and 12 teachers) to participate in the evaluation of several readings of texts 01, 02 and RoiGourmand. Most readings have been judged by three raters, while 78 were evaluated by more than 16 raters during listening sessions in lecture rooms.

Fig. 2 gives distributions of these ratings (averaged for each reading) for each subjective dimension as a function of grade (all texts combined). Note that the EXP and SMT of children are still immature at grade 6, EXP not really improving after grade 4. The average ratings are moderately correlated:

- PHR is rather correlated with PAC (0.85), SMT (.79) and EXP (0.77)
- PAC is correlated with SMT (.81) and EXP (0.68)
- EXP is the least correlated with others, in particular SMT (0.59)

Fig. 3 gives the distribution of the inter-rater agreement at the item level per pair of raters (for each pair, we compute the Kappa coefficient on the set of readings they have both rated), distinguishing between the level of raters' expertise. As expected, teachers and experts are more consistent than students. PAC has significantly better reliability than other dimensions. It is particularly high for teachers.

In the following, we will consider average ratings.

4. Objective assessment

The automatic assessment uses both verbal and prosodic features. The characterization of verbal performance (see Section 4.2) distinguishes between correct versus incorrect pronunciations of licit words as well as repetitions and insertions. A word is considered as being spelled out as soon as one syllable is produced. Prosodic features (see Section 4.3) are computed by comparing the reader's prosodic patterns to multiple adult readings.

4.1. Text-to-speech alignment

All text-to-speech alignments performed at the phone- and word-level have been hand-checked. The acoustic model, the pronunciation dictionary, and tri-gram language model of a GMM/HMM model – trained using Kaldi (Povey et al., 2011) – have been constantly updated as aligned data have been hand-checked. Our labeling policy consists in building an extended lexicon with correct and incorrect pronunciations of the text words as well out-of-text words (see an example in Fig. 4). The tri-gram model is thus also disfluent: it incorporates text-specific syntactic organization of false starts, repetitions, and incorrect pronunciations. These disfluent lexicons and tri-gram models are trained on the children's readings (see Godde et al., 2017, for the first mention of this forced-alignment policy). The automatic aligner used in Section 6 uses the same policy.

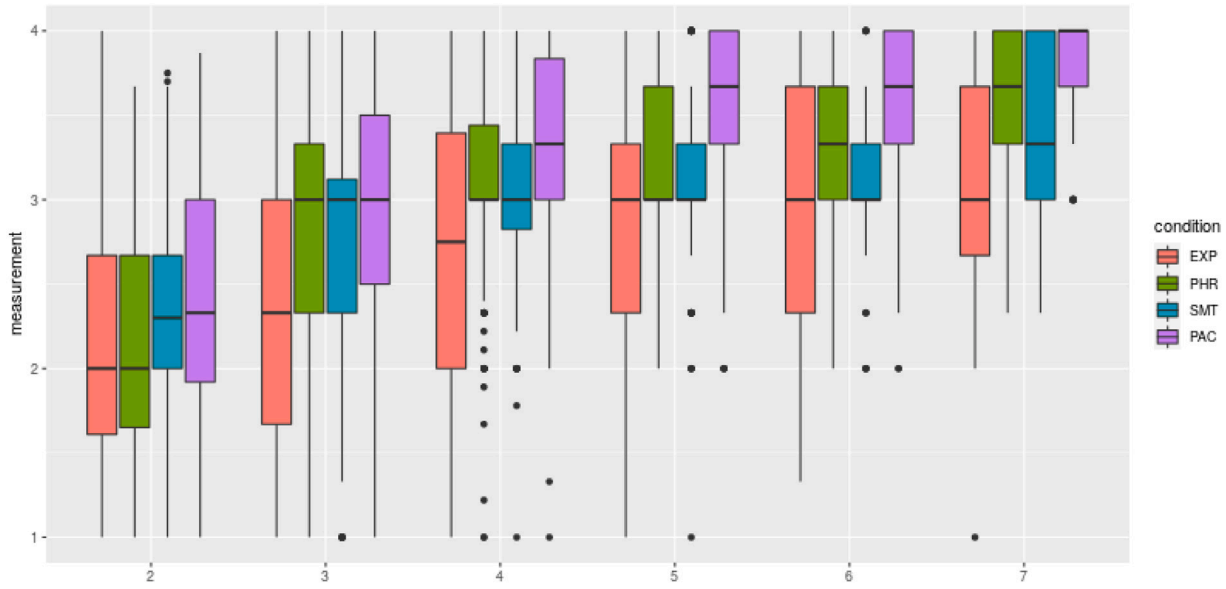

Fig. 2. Distributions of scores for each MDFS subjective dimension as a function of grade (all texts taken together). PAC, SMT, PHR and EXP respectively stand for pace, smoothness, phrasing and expressivity.

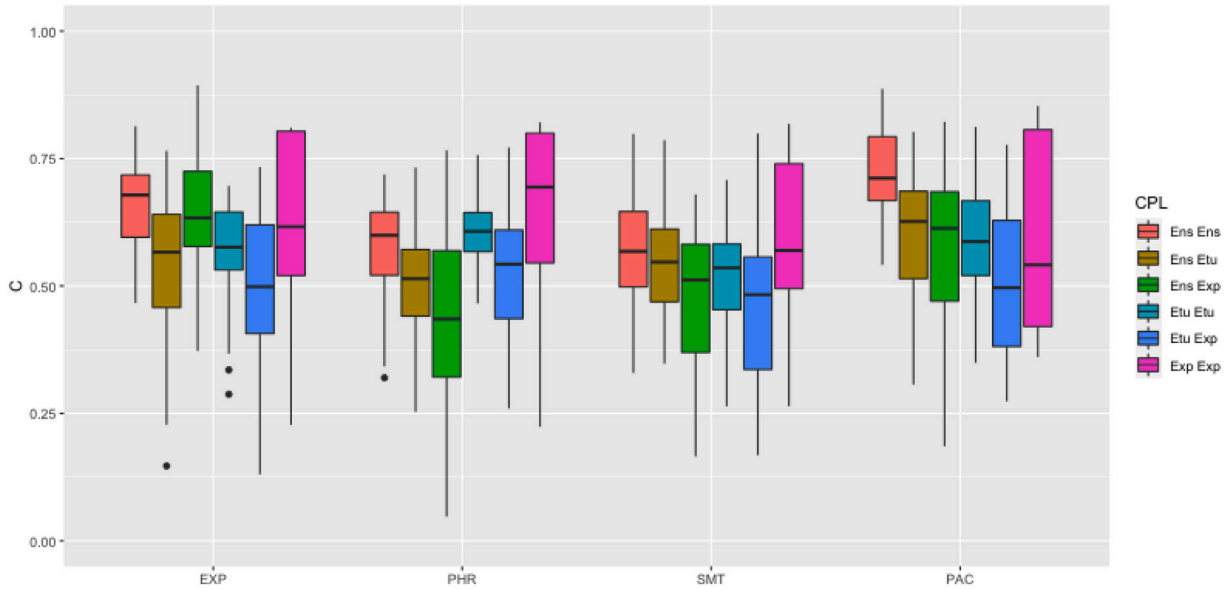

Fig. 3. For each subjective dimension, inter-rater reliability at the item level for the three different level of raters' expertise (Exp = experts, Ens = teachers, Etu = students). Within-group vs. inter-group Kappa coefficients are given.

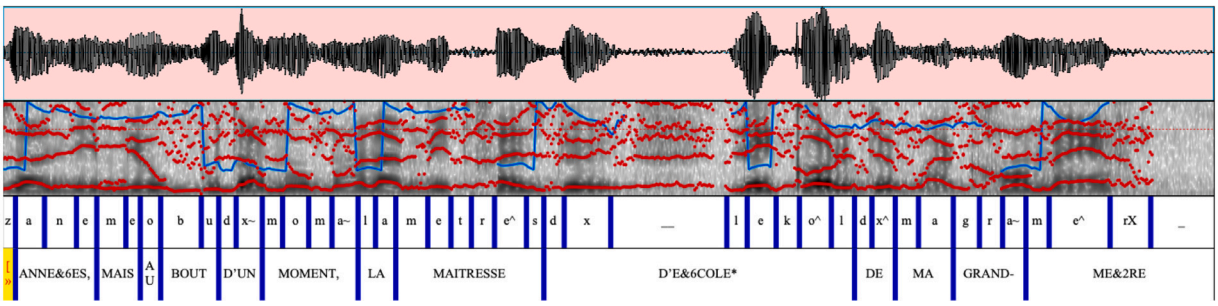

Fig. 4. Phone- and word-level alignment for Text A and a child reading. The noun complement ''De l'école'' (school teacher) was substituted with the compound ''d'école'' (teacher of the school), considered as mispronounced (the word is suffixed by the star symbol (*). Note also the long internal pause (__) that signals that the child is conscious of this error and has been hesitating.

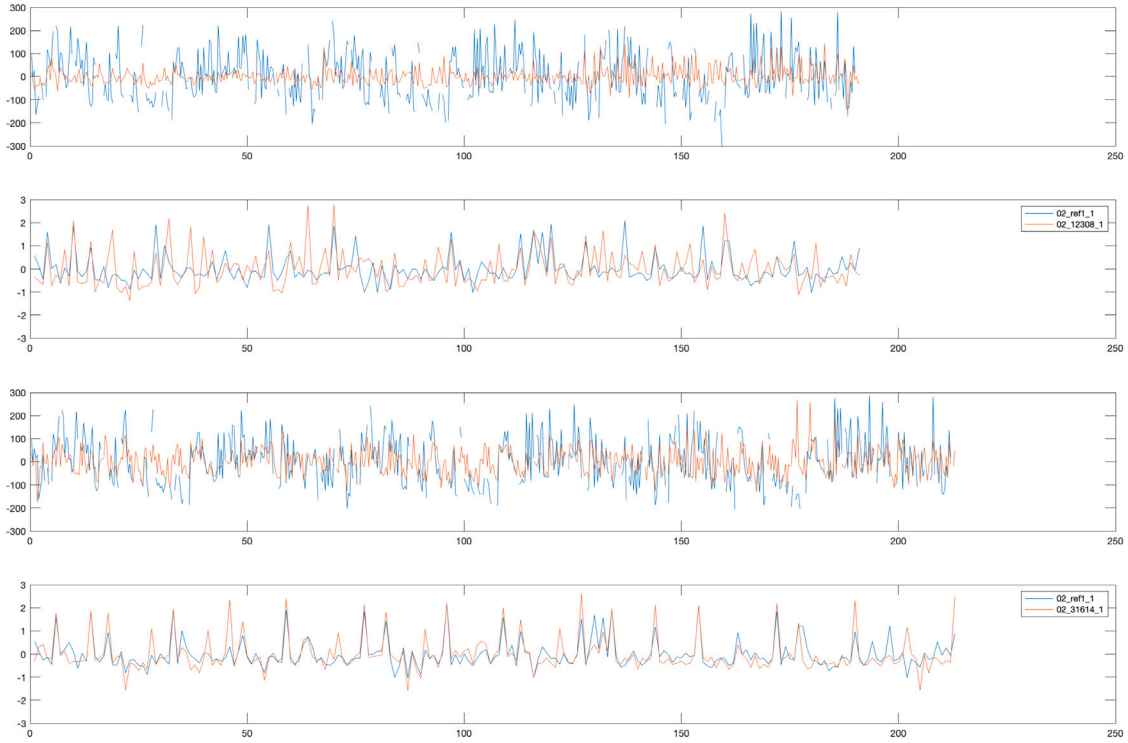

Fig. 5. Alignment of centered F0 (expressed in cents) and COE (syllabic stretching, see 4.3.1 for the z-score computation) of readings of a bad (two top tracings) vs. good (two bottom tracings) reader (red curves) against an adult reading (blue curves) for Text 02. Syllable counts of words correctly spelled are given in abscissa: the bad reader produces fewer correct syllables than the good reader (190 vs. 213). Contours respectively aggregate 3 F0 values and one COE value per syllable. Bad readers are often characterized by flat F0 and poor alignment of pauses (large COE values that are out-of-sync) while good readers have larger and relevant F0 excursions and better align with adult rhythmical contours. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Data collection. The number of correct vs. incorrect pronunciations correspond to the number of entries in the pronunciation dictionaries of each text (each entry being used by at least one reader). Note that correct pronunciations include all the false starts till one syllable. The average number of pronunciations per word given in parenthesis somehow relates to the text difficulty: texts 01 and 02 are clearly more difficult to decode than the two others. Mean and standard deviations of individual reading lengths are given in parenthesis in the "Speech duration" column. The two last columns give the actual number of young and adult readers for each text and in total: some readers have read several texts, and sometimes several times (as in pre- vs. post-evaluation).

Text	Nb. words	Nb. correct pronunciations	Nb. incorrect pronunciations	Nb. readings	Speech duration	Nb. children	Nb. adults
01	159	905 (5.69)	1018 (6.40)	421	10 h 18′(1′23″ ± 0′35″)	330	21
02	155	693 (4.47)	668 (4.31)	412	9 h 22′(1′21″ ± 0′29″)	316	21
A	271	583 (2.15)	288 (1.06)	134	3 h 49′(1′42″ ± 0′26″)	69	21
RoiGourmand	421	1184 (2.81)	1316 (3.13)	96	7 h 25′(4′38″ ± 2′14″)	90	21
Total				1063	30 h 54′	442	42

4.2. Characterization of verbal performance

We complemented the classical NCW with several counts that may separate decoding difficulty from speed:

- Number of incorrect words pronounced per minute (NIW)
- Number of repetitions per minute (NR)
- Number of intra-word silences per minute (NIWS)
- Number of vocalic nuclei per minute (NV)

Table 1 lists the number of correct and incorrect spelling variants used by at least one pupil for each text. When considering the average number of correct versus incorrect pronunciation variants per word, Text 01 is more difficult to decode. Despite the careful choice of words in Text 02, this simulated dialog is far more difficult to decode than the two last ones. These difficulties could be due to the type of text (i.e., a dialog) unfamiliar to the youngest readers. They hesitated and repeated a lot because of the unfamiliar ping-pong structure of the text.

4.3. Characterization of prosodic patterns

While the verbal content is imposed by the text, the variability of acceptable prosodic patterns is quite large and depends on numerous factors such as idiosyncratic breathing patterns, text comprehension, and interpretation, expressive strategies, etc. There is no "gold standard" prosody that makes one text rendering more relevant or likable than others.

Hirst et al. (1998) proposed to evaluate the predicted prosody of text-to-speech (TTS) systems by comparing the generated prosodic patterns to several natural references. After phonetic alignment and for each prosodic parameter (segmental duration, fundamental frequency (F0) and intensity), they considered the root-mean-square (RMS) distance between the value measured for the synthetic version and the most similar version of the natural recordings as a dissimilarity rating between the natural and the synthetic versions. This evaluation framework does not however enable the comparison of different prosodic renderings since natural references are not positioned in a unique global latent space.

Truong et al. (2018) also proposed an automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours built using k-means clustering of native readings. Experiments were conducted on isolated English words spoken by Japanese learners of English. Among all the utterances in the corpus, 910 were prosodically assessed by two native American English teachers who had to evaluate the prosody of the speakers on a scale of 1 to 5, corresponding to categories ranging from "very poor" to "excellent". They obtained a rather modest final subjective–objective score correlation of 0.3. the evaluation is here prescriptive: Truong et al. (2018) focus on whether or not the speaker positioned the stress on the appropriate syllable. Native reference utterances for each word present in the dataset were taken from online English dictionaries pronounced by speakers with various English accents: the 3 clusters for F0 and 2 clusters for intensity contours are likely to encode pronunciation variants.

Similar to what we proposed for the analysis of social gaze patterns (Bailly et al., 2010), we used MultiDimensional Scaling (MDS) to first map prosodic patterns of reference stimuli to points in a low n-dimensional space (n = 3 here) and then project children's prosodic patterns that will be finally characterized by a few loading factors.

4.3.1. Building prosodic spaces

We, therefore, processed the readings of 21 adults for each text. They were instructed that the readings will be used as reference patterns for children but no further constraints on parsing, emphasis, or expressiveness were given, so as to hopefully cover a large variety of styles (including motheries) and prosodic shapes on each word of the target text. After lexical and phonetic alignment (see typical alignments for bad and good readers in 5), we computed two inter-reference distances: one for the melodic pattern (F0) and one for syllabic stretching (COE). Only correct pronunciations are considered: we only compute cumulated distances between features of the last occurrence of each word correctly pronounced by both readers. These cumulated distances are then divided by the number of syllables considered. Note that unvoiced nuclei are discarded for the computation of F0 distances. A vocalic nucleus is considered unvoiced if less than 75% of its frames are unvoiced. F0 values (expressed in cents) are sampled at three positions within each syllabic nucleus (10%, 50%, and 90%). COE is expressed as a z-scored value. It equals the deviation between the syllabic duration (+ its optional following pause) and one expected duration, computed as a function of contributing segments (see Barbosa and Bailly, 1994, for more details). We used the same z-score model for all speakers, discarding idiosyncratic variations.

Before cumulating distances between prosodic features, we subtract the mean values of the prosodic contours: differences between speakers' registers and average reading speeds are thus suppressed. Note that average reading speed is already part of the verbal predictors (feature NV).

Once the 21 × 21 matrix of adult inter-reference distances is obtained, it is symmetrized and an MDS with n = 3 factors is performed for each feature (F0 and COE).

Note that these prosodic spaces are text-dependent: each text has its F0 and COE MDS factors, that somehow describe the licit modes of variations of the prosodic patterns for that particular text.

5. Inferring prosodic features of children's readings

The positioning of children's readings in two prosodic spaces (F0 and COE) is obtained (a) by aligning and computing cumulated distances with the 21 references; (b) projecting each distance vector onto the corresponding reference MDS space. Fig. 6 displays the positioning of the 21 references and the children's readings in the first MDS factorial planes for F0 and COE for Text 01. Figs. 7, 8 and 9 give the prosodic spaces from the other three texts. For each text, we indicated good, poor, and medium readers, according to subjective scores, with a colored ellipsis.

Table 2
Comparing the prediction of subjective ratings using two sets of objective features: verbal-only vs. verbal+prosody. Correlations (Pseudo R2 MacFadden Coefficient (McFadden, 1987)) and the significant predictors of the simplified model (keeping contributions with $p < 0.05$) are given. COE_{1-3} and $F0_{1-3}$ are the MDS loading factors of the prosodic spaces. These prosodic features mostly improve the prediction of expressivity (EXP) and phrasing (PHR). As expected, F0 and COE features are mainly used for EXP and PHR respectively.

Text	Dim	verbal-only		verbal+prosody	
01	EXP	.62	$1 + NV$	.78	$1 + F0_1 + NCW + COE_1$
	PHR	.77	$1 + NCW$	.83	$1 + NCW + COE_1$
	SMT	.74	$1 + NCW + NIW$	.75	$1 + NCW + NIW$
	PAC	.81	$1 + NIW + NV$	.83	$1 + NV + NIW$
02	EXP	.59	$1 + NV$	.78	$1 + COE_1 + F0_1 + NV + F0_2$
	PHR	.77	$1 + NV + NIW$	.87	$1 + NCW + COE_1$
	SMT	.82	$1 + NIW + NV$	.83	$1 + NV + NIW$
	PAC	.86	$1 + NV$	.88	$1 + NV + F0_1$
A	EXP	.62	$1 + NCW$	.81	$1 + NCW + COE_1 + F0_1$
	PHR	.57	$1 + NCW$	.77	$1 + COE_1 + COE_2$
	SMT	.77	$1 + NCW + NV$	.81	$1 + NCW + NV$
	PAC	.79	$1 + NCW$	.83	$1 + NCW$
RoiGourmand	EXP	.83	$1 + NV$	.90	$1 + NV + COE_1$
	PHR	.78	$1 + NV$	.82	$1 + NCW + COE_1$
	SMT	.74	$1 + NCW$	.79	$1 + COE_2$
	PAC	.76	$1 + NV$	.80	$1 + NV$

We added a virtual reader – named ref0 and indicated by a blue square in Figs. 6, 7, 8 and, 9 – with flat prosodic contours (F0 and COE values for all syllables equal to 0.0) as an additional reference, helping to locate children's performance. As expected, good reader clusters are closer to the reference adult readers, while poor readers are further from reference readers and closer to ref0.

Note that the COE projections of children cover a much larger portion of the adult working space than f0 projections do. This is mostly due to the prominent use of melody by adults to encode stylistic variations.

5.1. Prediction of subjective ratings

We compared text-specific predictions of mean subjective ratings for each pupil via a simple multi-linear regression of verbal-only versus verbal + prosody features. Predictions of the latter model for Text 02 are given in Fig. 10. Results considering all available data are given in Table 2. The respective simplified models (only considering significant contributions with $p < .05$) for each text and each subjective dimension are listed. Contributing features are given in decreasing order of their contributions. We computed pseudo R2 MacFadden (comparing Log-likelihoods of the simplified vs. full model) coefficients.

Prosodic features mostly benefit the prediction of expressivity (EXP) and phrasing (PHR). As expected, F0 and COE features are mainly used for EXP and PHR respectively. Note however that the first COE loading factor gives significant contributions to the prediction of EXP for all texts. Note that this is congruent with the cross-links between phrasing and expressivity already observed by Cowie et al. (2002) and Benjamin et al. (2013).

The large improvement of the prediction of PHR and EXP, which have comparable correlation levels with SMT and PAC once prosody is taken into account, confirms the efficiency of the proposed framework for characterizing prosodic patterns with multiple references.

5.2. Text dependency

Fig. 11 compares average human (blue) versus predicted scores (red) for the readings of Texts 01 and 02 performed by the same 282 pupils. These comparisons are carried out for each subjective dimension and the overall score (left to right). Please note that readings of text 02 were rated by far fewer human assessors than Text 01. The later text

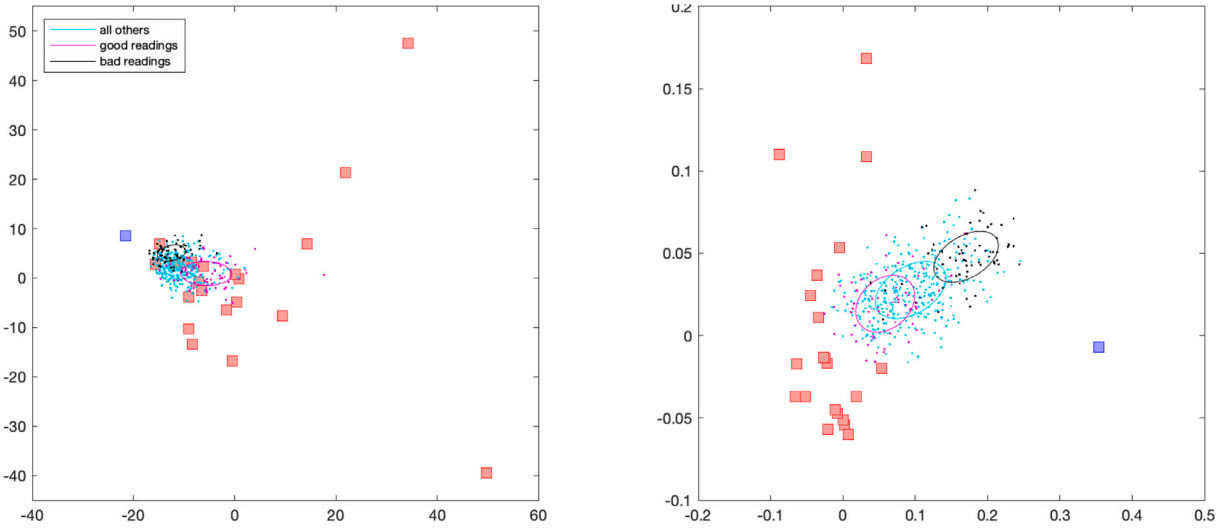

Fig. 6. Prosodic spaces calculated for Text 01. Adult readings are displayed with red squares while children are displayed with dots. Dispersion ellipsis sample children according to certain subjective performances given in the legend: bad, good and intermediate readers are respectively displayed in black, magenta and cyan. Bad readers are close to the artificial – flat and monotonous – ref0, displayed with a blue square. First factorial plans are displayed for F0 (left) and COE (right) spaces. Scales reflect distances between readings expressed in cents for F0 and syllabic stretching factor for COE. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

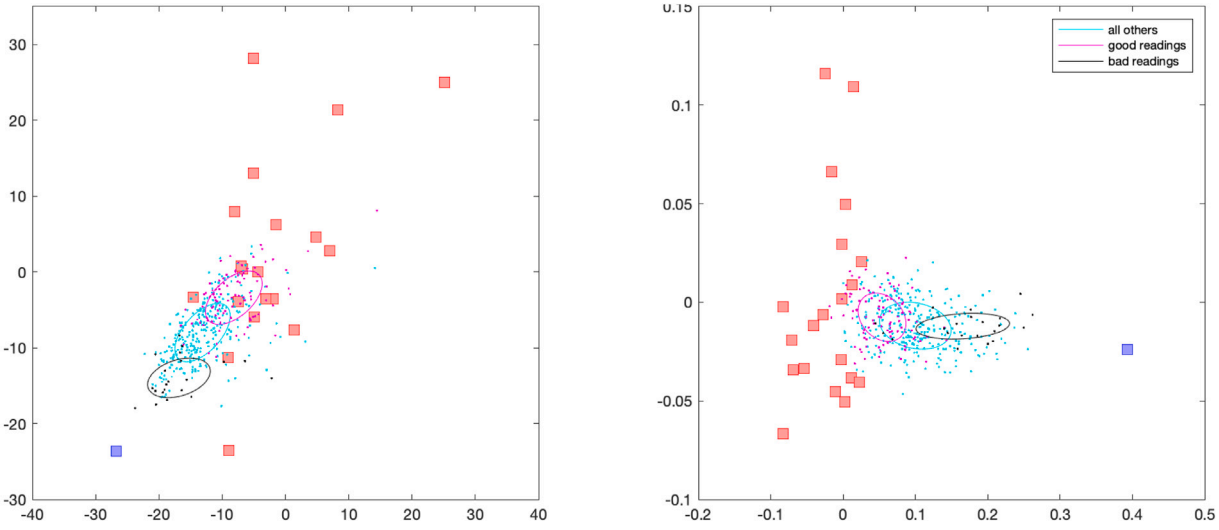

Fig. 7. Prosodic spaces calculated for text 02. Same conventions as Fig. 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

indeed served as a benchmark for studying the impact of education and training on scoring performance. Predicted scores are far more consistent than human ratings: correlations between predicted scores for 01 and 02 exceed 0.94 except for EXP which still reaches 0.88. Human scores are significantly less correlated. The lowest correlation between average human scores of the two texts is obtained for smoothness (SMT). A plausible explanation is that dialogs of text 02 are easier to process than the narrative style of text 01 and entail fewer disfluencies. Note that a general limitation of human assessment is that human raters just listen to the first minute of children's performances, while prediction is performed on the entire reading. Although the current prediction of expressivity seems to be moderately text-dependent, the high correlation values between predictions advocate for the robustness of automatic scoring and its practical use in evaluation sessions.

Table 3
Automatic assessment: means (± standard deviations) of cross-correlations between mean subjective (ground-truth data) and predicted scores for train and test partitions. Only results for text 01 and 02 are shown here because of insufficient training material in each partition for the other text readings.

Text	Dataset	EXP	PHR	SMT	PAC
01	Train	0.78 ± 0.005	0.83 ± 0.004	0.75 ± 0.005	0.83 ± 0.006
	Test	0.76 ± 0.050	0.79 ± 0.034	0.69 ± 0.048	0.75 ± 0.060
02	Train	0.78 ± 0.006	0.87 ± 0.004	0.83 ± 0.005	0.88 ± 0.004
	Test	0.74 ± 0.081	0.85 ± 0.048	0.81 ± 0.048	0.86 ± 0.054

6. Automatic assessment

We further assessed the automatic estimation of subjective ratings from raw speech signals without any manual intervention, in the following way:

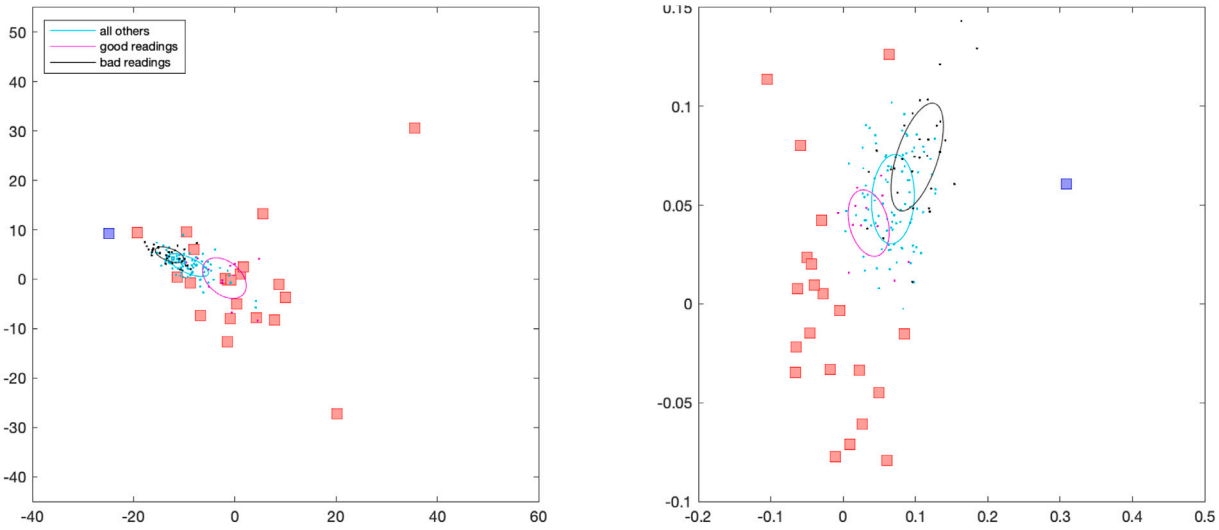

Fig. 8. Prosodic spaces calculated for text A. Same conventions as Fig. 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

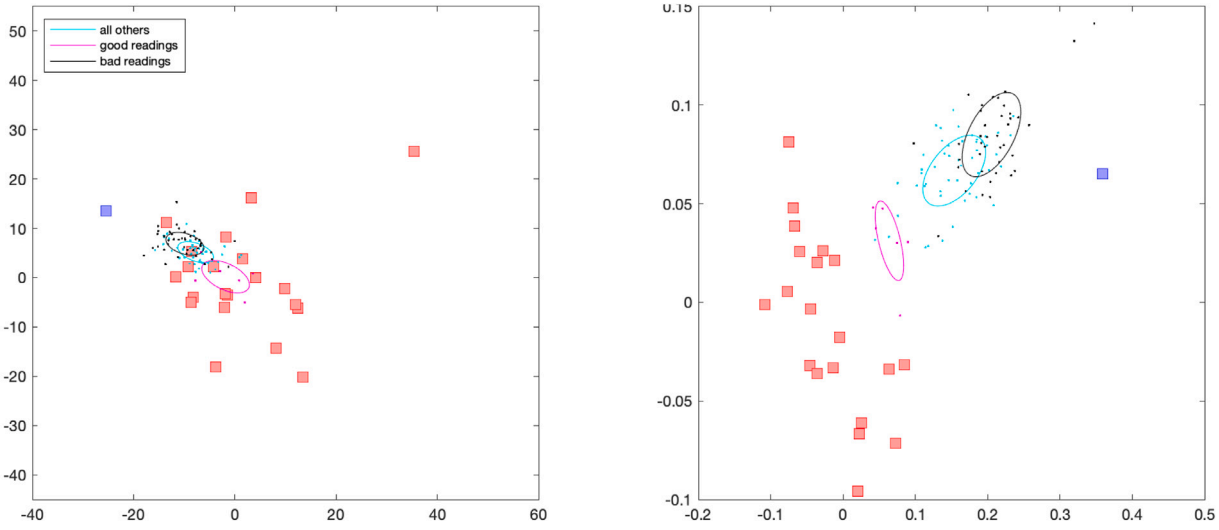

Fig. 9. Prosodic spaces calculated for text RoiGourmand. Same conventions as Fig. 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

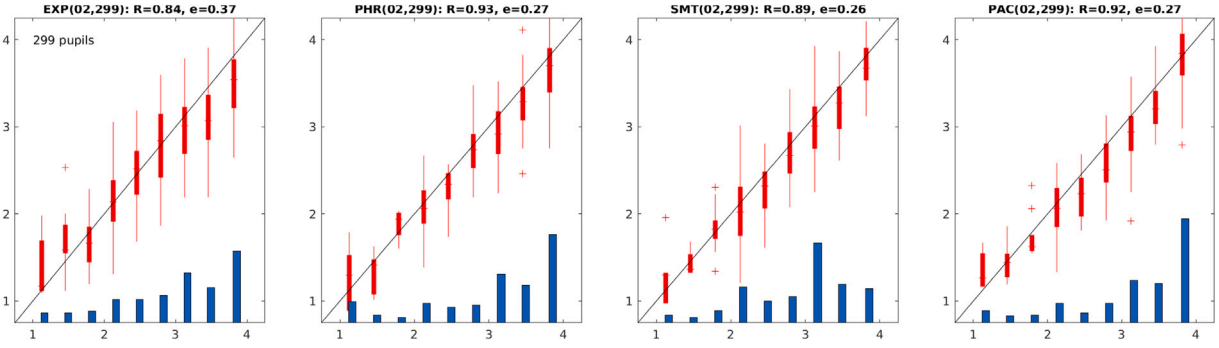

Fig. 10. For text 02, subjective (abscissa) vs. predicted (ordinate) ratings using verbal + prosody features for each subjective dimension: EXP, PHR, SMT and PAC (left to right). Correlations and mean RMS errors are given in the title of each caption. Empirical distributions of subjective ratings are superimposed in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

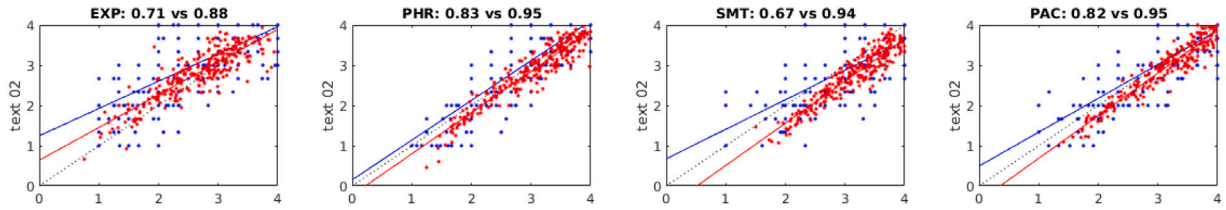

Fig. 11. Text dependency. Comparing average human (blue) vs. predicted scores (red) for the readings of texts 01 and 02 performed by the same 282 pupils. These comparisons are performed for each subjective dimension (left to right). Captions give the correlations between average scores given by humans for the two texts vs. those predicted by the regression models trained on the same data. Humans tend to overrate readings of Text 02, unlike the predictive models. Predicted scores are far more consistent than human ratings. This is particularly salient for SMT. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

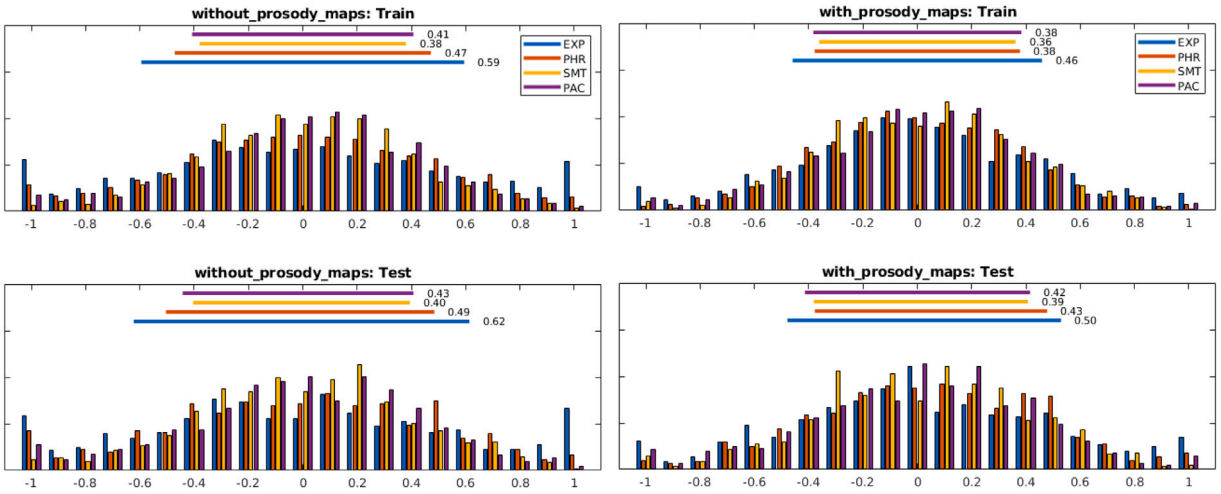

Fig. 12. Automatic assessment: distribution of prediction errors for training (top) vs. test (bottom) partitions, without (left) vs. with (right) taking prosodic features into account distributions for the 4 subjective dimensions are superimposed. Standard deviations are also superimposed with horizontal bars. As expected, significant reduction of variance is achieved for phrasing (PHR) and expressivity (EXP) when taking prosody into account.

Cross-validation we partition the readings of each text into 10 balanced subsets. Each partition is then used as test material while the remaining readings are used to train the resources for aligning (acoustic, lexical and syntactic models) oral readings and estimating multilinear regressors for mean subjective ratings

F0 estimation while F0 patterns of adult reference readings are carefully checked by hand, the Praat (Boersma, 2001) F0 extractor with a large F0 range (150 Hz–450 Hz) is used for children readings, with no further hand correction.

Automatic alignment We trained one GMM/HMM triphone acoustic model (TDNN front-end did not result in competitive results) and set up a pronunciation dictionary (with correct and incorrect entries) from all training data of the partition, and 4 text-specific trigram models per partition using Kaldi (Povey et al., 2011). We used the standard configuration (MFCC, VTLN) except that the energy term is discarded.

Multilinear regression Verbal and prosodic features (computed using text-specific prosodic spaces) are used to estimate subjective ratings of the training material following the procedure described in Section 5.

Fig. 12 displays the distributions of errors between predicted and subjective ratings for the training versus test data, compiled for all texts. The standard deviations of prediction errors for the test data stay below 0.5 points when models use text-specific prosodic features (COE and F0). When discarding prosodic features, the prediction of phrasing (PHR) and expressivity (EXP) is significantly degraded.

Table 3 finally gives the mean correlations between ground-truth subjective and automatically predicted scores for the training and test partitions. Only results obtained on texts 01 and 02 are given since other texts lack sufficient readings to provide enough data in each partition. These numbers should be related to those obtained by considering hand-labeled data with the `verbal+prosody` model (shown Table 2). Automatic alignment for estimating regression models and inferring predicted scores has almost no impact on training data. For test material, the degradation is close to 7%. The strongest degradations are observed for test SMT and PAC scores of text 01, due to shortcomings of the tri-gram model that result in an improper estimation of the verbal features (NIW, NCW). Since the regression models for text 02 rely less on verbal features than those for text 01 (see column 6 of Table 2), they are more robust to verbal misalignment.

7. Comments

Prosodic spaces Unlike the assignment of correct versus incorrect pronunciations, the projection of prosodic patterns in the adult prosodic spaces avoids a strict prescriptive evaluation that would distinguish between licit and illicit pitch accents, boundary tones or contours (e.g., such as the phrasal break detection and prominent word detection used by Sabu and Rao, 2018). For now, after alignment, the prosodic contours of all syllables equally contribute to the prosodic distances between children's performances and adults' anchor points. Weights could be added to give more importance to some phonological landmarks, in particular when some landmarks are highlighted in the text for training purposes using explicit instructions or implicit augmentation (such as the Karaoke technique proposed by Bailly and Barbour, 2011; Godde et al., 2017).

Note also that other prosodic dimensions – such as intensity or spectral tilt – may have been taken into account. Their strong dependence on recording conditions and phonetic support and their weak

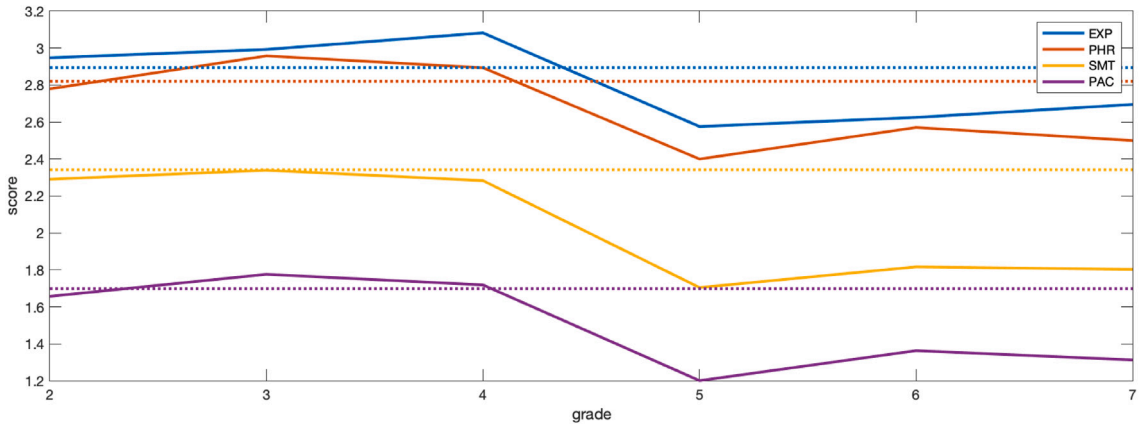

Fig. 13. Class level as an additional predictor of the verbal+prosody regression model. Intercepts of the class-independent model are shown in dotted lines. Because of the rise of performance expectations by raters, scores are lowered by about 0.4 points (0.6 for SMT!) when children are over the 4th grade.

contribution to final accentuation in French (Nemoto, 2011) led us to ignore them in this work.

Text material As evidenced in Fig. 2, text complexity should be adapted to the reader's developmental level to avoid saturation effects. We have shown Godde et al. (2020a) that texts 01 and 02 cover a large spectrum of young readers from 2nd to 5th grade. More complex texts with lexical content adapted to teenagers should be used for readers from 6th grade and above. As shown in Table 2, prosodic predictors are text-dependent. In that respect, F0 loading factors contribute far more to the evaluation of perceptual dimensions for text 02 than the other texts, since text 02, a dialog, was introduced to probe expressivity.

The age bias The predictive models are blind to the children's age and sex. Although instructed not to do so, human ratings are unconsciously biased by the expected performance of children given their estimated age. We were not allowed to collect age but had access to class level. When adding the class level in the model's predictors, raters seem to raise their expectations: with equal verbal and prosody performance, scores of children over the 4th grade are lower by about 0.4 (see Fig. 13). This is true for all subjective dimensions. This bias improves correlations between subjective means and predictions by a few percent. For texts 01 and 02, the correlations for the test data in Table 2 (Column 5) respectively raise to 0.79, 0.84, 0.77, 0.83 and 0.79, 0.88, 0.84, 0.88 for EXP, PHR, SMT and PAC. Since Proença et al. (2017) screened primary school children aged 6 to 10 years, they did not observe this phenomenon.

And adults? We processed the 21 reference adult readings for each text using the text-specific models trained on the children's readings, to verify that these strong extrapolation tasks give reasonable results. Table 4 gives mean and standard deviations of these predicted ratings. All mean estimations of the four subjective dimensions lie in the interval $[4.01 - 4.8]$ with predicted scores for texts 01 and 02 having the lowest variances. As expected, adults are rated as super fluent, on these texts that are calibrated for children (as observed in Godde et al., 2020a, text complexity should be adapted to the reading age) and with models that do not raise their expectations together with such an increase in reading abilities. A compensation scheme that would decrease performance by almost one point is however quite compatible with the 0.4 point downgrading observed for children from 5th to 7th grade.

Subjective evaluation Inter-rater agreement is of course a major issue. We observed here that two-by-two inter-rater agreement is low, both in and between most of the groups. Even with precise training and protocol, there are still differences in the assessment of young readers. Fig. 3 showed that inter-rater agreement is globally better for raters familiar with children's readings. It is also to be noticed that inter-rater agreement is slightly better for pace. Pace is indeed the less subjective dimension to assess. Variability in the expectation of the raters could explain a part of the differences. Moreover, subjective evaluation relies on a discrete scale: raters have to choose between 2 or 3, without the possibility to give a 2.5. This also could lead to low inter-rater agreement. Considering this, having multiple raters is an important issue to have scores as precise as possible to define the models. Once the model is defined for a text, the automatic assessment enables us to get rid of this inter-rater variability and gives a precise score for each reading. Multiple assessments of the same reader can then be easily compared.

Table 4
Predicting subjective ratings of adult readers from models trained on pupils. As expected, all scores are above 4 for these texts aimed at children.

Text	EXP	PHR	SMT	PAC
01	4.30 ± 0.99	4.27 ± 0.51	4.10 ± 0.45	4.12 ± 0.49
02	4.66 ± 0.98	4.72 ± 0.50	4.01 ± 0.43	4.65 ± 0.56
A	4.35 ± 1.35	4.56 ± 1.18	4.80 ± 1.05	4.35 ± 0.86
RoiGourmand	4.40 ± 1.17	4.56 ± 0.94	4.71 ± 1.55	4.64 ± 1.40

8 Conclusions

We have proposed an innovative framework for taking into account the multiplicity of licit prosodic patterns of oral readings of various texts by adult and young readers. We have shown that verbal and prosodic features, computed from the speech signals, can be combined to robustly predict perceptual dimensions that trace the acquisition of reading fluency.

Our main contribution concerns the use of multiple readings for computing prosodic features. Results of the automatic assessment show that the estimation of verbal cues is rather sensitive to the amount of training material and can largely be improved. Recent advances in transformer-based phone recognition (Gelin et al., 2021) can surely better cope with disfluencies and reading mistakes.

Similar regressions can be performed for predicting other dimensions such as text comprehension, which is known to be mutually influenced by prosody in the course of the development of reading fluency (Miller and Schwanenflugel, 2008).

The long-term goal of this work is to include this automatic assessment in a computer-assisted training curriculum for reading fluency, both in the classroom and in autonomy at home. This curriculum goes through the training of elementary skills such as the expansion of the visual attention span, phonological awareness, word recognition... as well as reading of pseudo-words and isolated sentences. But the ultimate goal is to improve fluency and online comprehension of whole

texts. Being able to regularly evaluate the progress of the children made on the acquisition of each reading skill and couple this evaluation with the training is a prerequisite for an efficient training program.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The protocol has been approved by the local ethical committee (CERNI-Avis-2018-03-06-2). This work has been financed by the CDC-founded e-FRAN Fluence project and MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). We warmly thank all pupils and teachers for their hospitality to the experimenters, Erika Godde and Manon Metz. Adults have been recorded by Anne-Claire Dugué and numerous text-to-phonetic alignments have been hand-checked by Margaux Manka.

Appendix. Annexes

The four texts used in this experiment are:

01 "Mon chat est entré dans ma chambre. Il avait l'air bizarre donc j'ai tout de suite compris qu'il avait fait une énorme bêtise. D'abord j'ai pensé au Poisson rouge. Je m'attendais à trouver le bocal vide. Mais mon Poisson était toujours là et il me regardait de ses grands yeux ronds. Ensuite je me suis souvenu que maman avait sorti des saucisses et les avait posées sur la table. Le chat les avait volées, c'était certain. Mais dans la cuisine, les saucisses étaient toujours là, prêtes à cuire, personne n'y avait touché. Le chat est alors allé vers mon cartable. Le cartable était ouvert, tout était par terre, mon exposé déchiré en tout petits bouts. Quand j'ai vu mon travail en morceaux, je me suis senti triste, j'avais envie de pleurer. Alors le chat s'est frotté sur ma jambe, il a miaulé, il était désolé. Finalement il m'a aidé et on a préparé ensemble un bel exposé sur l'intelligence des chats".

02 "Lucas arrive à l'école. Il retrouve son amie Lola.

« Salut Lola! Est-ce que tu as préparé ton exposé? demande Lucas.

– Oui. J'ai parlé de mon livre préféré, répond Lola. Et toi? Je suis sûre que tu as choisi les super-héros!

– Et non! Qui t'a dit que j'aimais les super-héros? s'étonne Lucas.

– Enzo, je crois. Tu l'as fait sur quoi alors ton exposé? poursuit Lola.

– Il se mêle toujours de tout celui-là, grogne Lucas.

– Allez, dis-moi! Quel sujet as-tu choisi finalement? reprend Lola. Les lions? La danse? La musique?

– C'est une histoire amusante, répond Lucas. Est-ce que je t'ai déjà parlé de mon chat?

– C'est celui qui fait toujours des bêtises? Je l'ai déjà vu chez toi, se souvient Lola.

– Oui, c'est bien lui. Hier il a mangé mon exposé sur les super-héros. Mais pour se faire pardonner, il m'a aidé à en faire un nouveau sur les chats! »"

A "Pourquoi les crayons à papier n'ont pas de jambes?

Ma grand-mère m'a raconté que lorsqu'elle était enfant, dans son école, les crayons à papier avaient des jambes. De toutes petites jambes qui leur permettaient de jouer en liberté. C'était très amusant... mais ce n'était pas pratique. En effet, ils passaient tout leur temps à courir dans la salle. Et lorsqu'un enfant cherchait son crayon à papier pour écrire des calculs sur sa feuille, il ne le trouvait jamais.

« Maitresse, disait-il, je ne peux pas faire mes calculs. Je ne trouve pas mon crayon à papier. Maitresse, je ne peux pas faire le dessin de ma poésie, mon crayon à papier est parti. »

Cela dura de nombreuses années. Mais au bout d'un moment, la maitresse d'école de ma grand-mère en a eu assez de cette agitation! Alors, un matin, elle a organisé avec ses élèves une chasse aux crayons à papier. Ils ont fermé les portes et capturé tous les crayons à papier avec leur filet à papillons.

Couic et couic... avec sa paire de ciseaux, la maitresse a coupé les petites jambes des fugitives. C'est depuis cette époque, paraît-il, que les crayons à papier n'ont plus de jambes. Parfois, ils essayent encore de s'échapper en tombant par terre et en rebondissant sur leur tête".

Le roi gourmand "Le roi est gourmand; ce qu'il aime le plus au monde, ce sont les figues. Il les aime tant qu'il promet la main de sa fille à qui lui apportera les meilleures figues du royaume. La nouvelle se répand, et parvient aux oreilles d'un brave paysan, qui, justement, cultive des figues avec l'aide de ses trois fils. Le brave paysan se décide:

« Tente ta chance, » dit-il à son fils aîné.

Le garçon approuve, prépare le panier de fruits et prend la route du château. En chemin, il rencontre une vieille femme... Celle-ci lui demande:

« Que portes-tu dans ton panier, jeune homme?

Le garçon pense que la vieille est trop curieuse et se mêle de ce qui ne la regarde pas:

– J'ai des crottes dans mon panier, répond-il grossièrement.

– Ah... Bien. Bonne chance, jeune homme. »

Le garçon arrive au château, annonce qu'il vient pour le concours. On l'introduit aussitôt dans la salle du trône.

« Approche, dit le roi. Fais-moi voir ce que tu apportes.

La figure gourmande, il ouvre lui-même le panier. Là, il sursaute de surprise et de colère:

–Des crottes! crie-t-il. Qu'on chasse ce maraud à grands coups de bâton! »

Les soldats obéissent. Couvert de meurtrissures, le garçon retourne chez lui, tout penaud. Racontant sa mésaventure, il se garde bien de parler de la vieille femme.

« Je n'y comprends rien, dit son père le paysan.

Alors, il décide que son deuxième fils ira à son tour au château, avec un nouveau panier.

– Choisis bien les fruits, » lui conseille-t-il.

Le deuxième fils s'en va. Lui aussi rencontre la vieille femme, qui lui demande ce qu'il transporte dans son panier. Lui aussi répond avec méchanceté qu'il transporte des crottes...

« Bonne chance, jeune homme. »

Il arrive donc au deuxième fils ce qui est arrivé au premier : ses figues sont transformées en crottes. Le roi ouvre le panier, se fâche... Il est, comme son frère, battu et chassé sans ménagement. À la maison, le père lève les bras au ciel en apprenant la nouvelle. Mais il est têtu :

« Tu vas tenter ta chance, » dit-il à son troisième fils.

Celui-ci aussitôt est d'accord. Le voilà donc, son panier au bras, sur la route du château, où il rencontre, on s'en doute, la vieille femme que ses frères ont déjà croisée.

« Que portes-tu dans ton panier, jeune homme?

– Je porte des figues au roi, madame, de belles figues, je crois. Voulez-vous en goûter une? »"

References

Álvarez-Cañizo, M., Martínez-García, C., Cuetos, F., Suárez-Coalla, P., 2020. Development of reading prosody in school-age Spanish children: a longitudinal study. J. Res. Read. 43 (1), 1–18.

Álvarez-Cañizo, M., Suárez-Coalla, P., Cuetos, F., 2018. Reading prosody development in Spanish children. Read. Writing 31 (1), 35–52.

Applegate, M.D., Applegate, A.J., Modla, V.B., 2009. "She's my best reader; she just can't comprehend": Studying the relationship between fluency and comprehension. Read. Teach. 62 (6), 512–521.

Ardoin, S.P., Morena, L.S., Binder, K.S., Foster, T.E., 2013. Examining the impact of feedback and repeated readings on oral reading fluency: Let's not forget prosody. Sch. Psych. Q. 28 (4), 391.

Bailly, G., Barbour, W.-S., 2011. Synchronous reading: learning french orthography by audiovisual training. In: Interspeech. pp. 1153–1156.

Bailly, G., Raidt, S., Elisei, F., 2010. Gaze, conversational agents and face-to-face communication. Speech Commun. 52 (6), 598–612.

Barbosa, P., Bailly, G., 1994. Characterisation of rhythmic patterns for text-to-speech synthesis. Speech Commun. 15 (1–2), 127–137.

Benjamin, R.G., Schwanenflugel, P.J., 2010. Text complexity and oral reading prosody in young readers. Read. Res. Q. 45 (4), 388–404.

Benjamin, R.G., Schwanenflugel, P.J., Meisinger, E.B., Groff, C., Kuhn, M.R., Steiner, L., 2013. A spectrographically grounded scale for evaluating reading expressiveness. Read. Res. Q. 48 (2), 105–133.

Blaauw, E., 1994. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. Speech Commun. 14 (4), 359–375.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot. Int. 5 (9), 341–345.

Bolanos, D., Cole, R.A., Ward, W.H., Tindal, G.A., Schwanenflugel, P.J., Kuhn, M.R., 2013. Automatic assessment of expressive oral reading. Speech Commun. 55 (2), 221–236.

Cowie, R., Douglas-Cowie, E., Wichmann, A., 2002. Prosodic characteristics of skilled reading: Fluency and expressiveness in 8—10-year-old readers. Lang. Speech 45 (1), 47–82.

Dowhower, S.L., 1991. Speaking of prosody: Fluency's unattended bedfellow. Theory Pract. 30 (3), 165–175.

Duong, M., Mostow, J., Sitaram, S., 2011. Two methods for assessing oral reading prosody. ACM Trans. Speech Lang. Process. 7 (4), 1–22.

Friot, B., 2007. Histoires Pressées. Milan.

Gelin, L., Pellegrini, T., Pinquier, J., Daniel, M., 2021. Simulating reading mistakes for child speech transformer-based phone recognition. In: Interspeech. pp. 3860–3864.

Godde, E., Bailly, G., Bosse, M.-L., 2022. Pausing and breathing while reading aloud: development from 2nd to 7th grade. Read. Writing 35, 1–27.

Godde, E., Bailly, G., Escudero, D., Bosse, M.-L., Gillet-Perret, E., 2017. Evaluation of reading performance of primary school children: Objective measurements vs. subjective ratings. In: International Workshop on Child Computer Interaction (WOCCI). Glasgow, UK.

Godde, E., Bosse, M.-L., Bailly, G., 2020a. Échelle multi-dimensionnelle de fluence: nouvel outil d'évaluation de la fluence en lecture prenant en compte la prosodie, étalonné du CE1 à la 5ème. L'année Psych. 19–43.

Godde, E., Bosse, M.-L., Bailly, G., 2020b. A review of reading prosody acquisition and development. Read. Writing 33, 399–426.

Haskins, T., Aleccia, V., 2014. Toward a reliable measure of prosody: an investigation of rater consistency. Int. J. Educ. Soc. Sci. 1 (5), 102–112.

Hirst, D., Rilliard, A., Aubergé, V., 1998. Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. In: The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.

Howell, P., Kadi-Hanifi, K., 1991. Comparison of prosodic properties between read and spontaneous speech material. Speech Commun. 10 (2), 163–169.

Kuhn, M.R., Schwanenflugel, P.J., Meisinger, E.B., 2010. Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. Read. Res. Q. 45 (2), 230–251.

Lefavrais, P., 1965. Description, definition et mesure de la dyslexie utilisation du test "l'alouette". Rev. Psych. Appl. 15 (2), 33–34.

McFadden, D., 1987. Regression-based specification tests for the multinomial logit model. J. Econometrics 34 (1–2), 63–82.

Meisinger, E.B., Bradley, B.A., Schwanenflugel, P.J., Kuhn, M.R., Morris, R.D., 2009. Myth and reality of the word caller: The relation between teacher nominations and prevalence among elementary school children. Sch. Psych. Q. 24 (3), 147.

Miller, J., Schwanenflugel, P.J., 2008. A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. Read. Res. Q. 43 (4), 336–354.

Moser, G.P., Sudweeks, R.R., Morrison, T.G., Wilcox, B., 2014. Reliability of ratings of children's expressive reading. Read. Psych. 35 (1), 58–79.

Nemoto, R., 2011. Large-Scale Acoustic and Prosodic Investigations of French (Ph.D. thesis). Université Paris–Sud — Faculté des Sciences d'Orsay.

Nicolson, A., Paliwal, K.K., 2019. Deep learning for minimum mean-square error approaches to speech enhancement. Speech Commun. 111, 44–55.

Pinnell, G.S., et al., 1995. Listening to Children Read Aloud: Data From NAEP's Integrated Reading Performance Record (IRPR) at Grade 4. ERIC.

Porte, E., Capek, J., 1998. Le Roi Gourmand. Epigones.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Catalog No.: CFP11SRW-USB, IEEE Signal Processing Society.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., Perdigão, F., 2017. Automatic evaluation of reading aloud performance in children. Speech Commun. 94, 1–14.

Rasinski, T.V., 2004. Assessing reading fluency. Pac. Resour. Educ. Learn..

Rasinski, T., Paige, D., Rupley, W., Young, C., 2019. Reading fluency: From theory to proof of concept, and from science to art. In: 69th Annual Literacy Research Association Meeting. Tampa, FL.

Sabu, K., Rao, P., 2018. Automatic assessment of children's oral reading using speech recognition and prosody modeling. CSI Trans. ICT 6 (2), 221–225.

Schwanenflugel, P.J., Hamilton, A.M., Kuhn, M.R., Wisenbaker, J.M., Stahl, S.A., 2004. Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers. J. Educ. Psych. 96 (1), 119.

Smith, G.S., Paige, D.D., 2019. A study of reliability across multiple raters when using the NAEP and MDFS rubrics to measure oral reading fluency. Read. Psych. 40 (1), 34–69.

Truong, Q.-T., Kato, T., Yamamoto, S., 2018. Automatic assessment of L2 english word prosody using weighted distances of F0 and Intensity Contours. In: Proc. Interspeech 2018. pp. 2186–2190.

Wolters, A.P., Kim, Y.-S.G., Szura, J.W., 2020. Is reading prosody related to reading comprehension? A meta-analysis. Sci. Stud. Read. 1–20.

Zorman, M., Lequette, C., Pouget, G., Devaux, M., Savin, H., 2008. Entraînement de la fluence de lecture pour les élèves de 6e en difficulté de lecture. In: ANAE, 96, Vol. 97. pp. 213–219.

Zutell, J., Rasinski, T.V., 1991. Training teachers to attend to their students' oral reading fluency. Theory Pract. 30 (3), 211–217.