Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 317–322
Marseille, 11–16 May 2020

Predicting Multidimensional Subjective Ratings of Children' Readings from the Speech Signals for the Automatic Assessment of Fluency

Gérard Bailly[1], Erika Godde[1,2], Anne-Laure Piat-Marchand[1] & Marie-Line Bosse[2]
[1]GIPSA-Lab, CNRS , Grenoble-INP & Univ. Grenoble-Alpes
[2]LPNC, CNRS & Univ. Grenoble-Alpes
{firstname.familyname}@gipsa-lab.fr, marie-line.bosse@univ-grenoble-alpes.fr

Abstract

The objective of this research is to estimate multidimensional subjective ratings of the reading performance of young readers from signal-based objective measures. We here combine linguistic features (number of correct words, repetitions, deletions, insertions uttered per minute ...) with phonetic features. Expressivity is particularly difficult to predict since there is no unique golden standard. We here propose a novel framework for performing such an estimation that exploits multiple references performed by adults and demonstrate its efficiency using recordings of 273 pupils.

Keywords: reading aloud, text-to-speech alignment, prosody, children voices

1. Introduction

Acquisition of fluency is a critical component of reading development. Reading aloud text necessitates the maturing and synchronization of multiple processes (Breznitz, 2006), from the visual decoding of letters, letters-to-sound mapping, lexical access, linguistic processing and comprehension, speech planning and articulation. While fluent readers are able to incrementally perform one-line text processing and utter its content with appropriate pacing, phrasing, and intonation, readings of young readers are characterized by dysfluencies, improper placements of pauses and flat intonation.

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

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

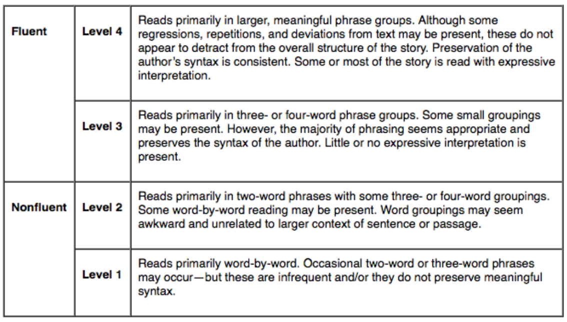
Grouping consists in successfully online grouping content words with their adjacent function words.

Phrasing consists in successfully pacing word grouping into meaningful linguistic units – in particular breathing and pausing at appropriate places.

Expressivity consists in successfully selecting adequate prosodic patterns from the online comprehension of the text, the situation and communicative intents.

Reading automaticity is considered to be acquired after the two first steps, i.e. during the second year of primary school.

In contrasts with our previous work (Godde et al., 2017) which was using a difficult text material (Lefavrais, 1967), we recorded children while reading aloud a single text with no traps in order to observe a large variety of phrasing and expressivity performance. We asked 3 experienced adult listeners to rate these readings according to the multidimensional fluency scale proposed by Zutell & Razinski (Zutell and Rasinski, 1991) and adapted for French. We here address the challenge of predicting these ratings from objective characterizations of the aloud readings.



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

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

2. State of the art

Reading fluency has long been defined as reading accurately and automatically: it is often still evaluated as the number of correct words pronounced per minute (CWPM). However, for a few decades, a new term appears in the key features of fluency : *reading prosody* (Miller and Schwanenflugel, 2008). Fluency is no longer a matter of accuracy and speed, but it takes into account the listener and the communicative aim of reading. One of the earliest publication including reading prosody in the definition of reading fluency comes from (Dowhower, 1991). She defined prosodic reading as "[...] the ability to read in expressive rhythmic and melodic patterns". In that respect, she proposed six relevant acoustic features of mature reading prosody: appropriate pausal intrusion, phrase segmentation and length, phrase-final lengthening, terminal intonation contours and stress. Since then, several authors have proposed to characterize children reading prosody with reference to mature realizations.

Cowie et al (Cowie et al., 2002) measured 40 different acoustic markers in the recordings of 8-10 years readers. They related these acoustic markers to subjective ratings of these recordings. It appears that the acoustic correlates of fluency and expressiveness are the one expected by the very definition of the terms. That is to say, fluency is mainly correlated to the basic temporal organization : pause duration, pause frequency, syllabic rate and pitch movement frequency. Expressiveness is mainly linked to pitch variation, i.e., pitch movement magnitude and duration and their variation from one sentence to another. It is to be noted that temporal organization and pitch variation are also intercorrelated.

Bolaños et al (Bolaños et al., 2013) complemented CWPM with numerous 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 in the text passage. They used Support Vector Machines (SVM) to predict NAEP-2 (fluent vs. non-fluent) and NAEP-4 ratings of 313 1st to 3rd grade students performed by two experts. They reported a machine-to-human Spearman's rank correlation of .86. But the NAEP (Danne et al., 2005) scale provide global ratings and do not distinguish between subjective dimensions. They claim that current speech technology may provide robust CWPM estimates but conclude that ”reading fluency scales have not (as yet) been grounded in research on reading prosody”. In fact, CWPM is computed against a golden standard of correct pronunciations that is rather easy to agree on (note however that this set may be quite large, see section 3.3.).

In contrast with the rather well-defined golden standard used to compute CWPM, the space of licit prosodic patterns is not so easy to define: good reading prosody depends on many factors such as the reader's specific breathing patterns, dialectal variation, interpretation of text content . . . .

In the following, we propose to calculate a prosodic space given readings of several mature readers.

3. Speech data

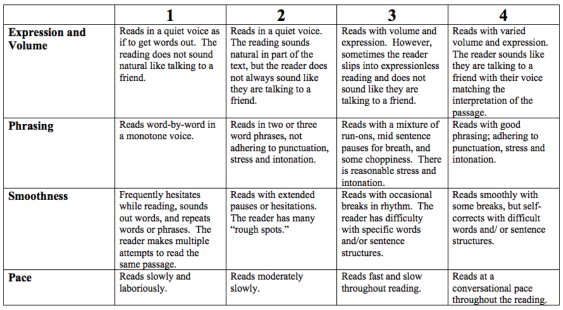
3.1. Recordings of aloud readings

As part of a larger study, we recorded 273 pupils from 2nd to 7th grade, aged 7 years 1 month to 13 years 9 months (mean = 10y2m) and 20 adults. The children were recorded in their schools, 2 primary schools and a middle school from Grenoble area, with the authorization of the schools directors and their parents. In terms of fluency, the children are representative of their grade-level, according to the fluency test Evaleo 6-15 (Launay, 2018). The adults were recorded on a voluntary basis in the lab, and were all assessed as expert readers by the assessors. During reading, we recorded their voice using a Schur Beta 53 microphone and a Berhinger MIC100 amplifier. The text used for the prediction is a 174-words narrative text written by the authors, with 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”.

3.2. Subjective Ratings

The subjective rating of the 273 recordings was performed by three assessors using the adapted Zutell & Rasinsky multidimensional fluency scale (see Figure 2) for French (Godde et al., submitted). They listen to the first minute of each recording to rate :

1. Pace (PAC) rates the speed of reading
2. Smoothness (SMT) rates the quality of on-line processing of words
3. Phrasing (PHR) rates the quality of word chunking and pause placement
4. Expression and volume (EXP) rates the reading expressivity



	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.

Figure 2: Multidimensional fluency rating scale (from (Zutell and Rasinski, 1991)).

Each dimension was given a score between 1 (no skills) and 4 (expert skills) according to the multidimensional rating scale, leading to a maximum cumulative total of 16. The assessors first accorded their ratings on 10 recordings, to clarify the ratings parameters. Then they independently rated the other recordings. The inter-rater agreement, calculated with Krippendorf's alpha (Hayes and Krippendorff, 2007) for ordinal data, is 0.96 for the total score. Regarding each parameter, the inter-rater agreement is 0.82 for expressivity and phrasing, 0.8 for smoothness and 0.81 for pace. Note that pace and smoothness were found easy to subjectively rate by the assessors, whereas expressivity and phrasing seemed more complex to assess, mostly because of intra- and inter-speaker variability of the licit prosodic patterns.

We then used the average of the rating of the 3 assessors as the subject's prosodic score. The children total scores range from 1 to 16 (11.29±2.23) while the adult ones range from 14.3 to 16 (15.42±1.45).

3.3. Objective characterization of verbal content

The audio signals were automatically aligned with a statistical model whose phonetic triphone models, pronunciation dictionary (with correct and incorrect pronunciations of words) and trigram model (capturing syntactic constraints on omission, repetetion . . .) were constantly updated using HTK and SLIRM toolkits once each alignment has been hand-corrected. While most speech recognizers consider mispronunciations and disfluencies differently from standard entries into the pronunciation dictionary, we treat correct, incorrect or incomplete words the same way in the pronunciation dictionary and the language model.

The labeling of words was thus performed with the following principles: (a) a new word is considered as initiated when at least one vocalic nucleus has been spelled; (b) a star is appended to any incorrect or incomplete word; (c) each dictionary entry begins with a phonated sound and syntactic and respiratory pauses are considered as part of the preceding word. The 273 readings of the 174 words of the text result in 1241 different correct vs. 1111 incorrect/incomplete pronunciations (see excerpt in table 1). Note that 67 correct vs. 15 incorrect/incomplete entries

Table 1: Excerpt of the incorrect pronunciation dictionary for the words "l'intelligence", "saucisse" or "finalement". These entries are suffixed by "*" and added to the licit entries. The n-gram model thus captures syntactic constraints on false starts, repetitions due to incorrect retrievals of pronunciations .... _ and __ respectively stand for syntactic vs respiratory pauses. Note that internal pauses may be also encountered.

nb	word	pronunciation
3	L'INTELLIGENCE*	le˜_
1	L'INTELLIGENCE*	leˆr__telizˆa˜s
1	L'INTELLIGENCE*	le˜telazˆizˆa˜s_
1	L'INTELLIGENCE*	le˜teliz_
2	L'INTELLIGENCE*	le˜telizˆa˜zˆ
1	L'INTELLIGENCE*	le˜tereˆs__
1	L'INTELLIGENCE*	le˜tezˆizˆi__izˆa˜s_
1	L'INTELLIGENCE*	li_teˆ__
1	L'INTELLIGENCE*	lxˆ_
1	FINALEMENT*	fe˜la_ma˜__
1	FINALEMENT*	fin_
1	FINALEMENT*	fina_
1	FINALEMENT*	final__
1	FINALEMENT*	finalm__
3	SAUCISSES*	sosiso˜
1	SAUCISSES*	susi_

comprise internal pauses – sometimes include air intakes. These internal pauses are mainly produced by Level 1 children with decoding problems.

3.4. Objective characterization of prosody

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

Hirst et al (Hirst et al., 1998) proposed to evaluate the predicted prosody of text-to-speech (TTS) systems by comparing it to several natural references. After phonetic alignment and for each prosodic parameter (segmental durations, fundamental frequency (F0) and intensity), they consider 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 innovative evaluation framework does not however enable the comparison of different prosodic renderings since natural references are not positioned in a unique global latent space. Similar to what we proposed for the analysis of social gaze patterns (Bailly et al., 2010), we used multidimensional scaling (MDS) to first represent prosodic patterns of reference stimuli as points in a n-dimensional space. Prosodic patterns of children' reading will be then represented by their projection onto this reference space.

We therefore processed the readings of 20 adults. They were instructed that the readings will be used as reference patterns for children but no further constraints on parsing, emphasis nor expressiveness were given, so that to hopefully cover a large variety of prosodic shapes on each word of the target text.

After phonetic and lexical alignment, we computed two inter-reference distances: one for F0 and one for syllabic stretching (COEFF). For all alignments (reference and test children readings), we do not consider incorrect spellings, omissions and repetitions, i.e. we only compute cumulated distances between features of the last occurrence of each word correctly spelled by both readers (see typical alignments for bad and good readers in Figure 4). F0 values (expressed in cents) are sampled at three positions within each syllabic nucleus (10%, 50% and 90%). COEFF is expressed as z-score coefficient (deviation between the syllabic duration (+ its optional following pause) and an expected duration, computed as a function of contributing segments (for more details see (Barbosa and Bailly, 1994)). We used the same z-score model for all speakers. Before cumulating distances between syllabic features, we subtract the mean values of the prosodic contours: differences between registers and average reading speeds are thus suppressed.

We added a virtual reader – named ref0 – with a flat F0 contour and a constant COEFF as an additional reference. Once the 21x21 matrix of adult inter-reference distances is obtained, it is symmetrised are an MDS with 3 factors is performed on each feature, namely F0 and COEFF. The positioning of children' readings in these 3D spaces is obtained (a) by aligning them and computing cumulated distances with the 21 references; (b) projecting this distance vector onto the reference 3D space. Figure 3 displays the positioning of the 21 references and the children readings in the first MDS factorial planes for F0 and COEFF. Please note that F0 projections of children readings are close to the flat ref0, while COEFF projections cover a much larger area of the maximal reference space: most first grade pupils are in the process of mastering fluency level 3.

4. Predicting subjective ratings

For each child's reading, we performed the alignment of the uttered words with the original text and computed the following features (number per minutes):

- CWPM: nb. of correct words per mn
- IWPM: nb. of omitted, incorrect and repeated words per mn
- VPM: nb. of vowels per mn

We 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 ratings using a Leave-One-Out procedure. The Spearman correlation coefficients (Scc) are respectively .66, .82, .86 and .85 for EXP, PHR, SMT and PAC (see top of Figure 5) with mean absolute errors equal to .50, .36, .27 and .33. All correlation coefficients are highly significant. As already evidenced (Godde et al., 2017), CWPM, IWPM and VPM are good predictors of PHR, SMT and PAC, but insufficient to accurately estimate EXP.

When the three first MDS loading factors of F0 and COEFF are added to the set of predictors, the prediction errors of EXP and PHR significantly improve and reach the same level of performance as for SMT and PAC: Scc are now respectively .85, .88, .87 and .88 for EXP, PHR, SMT and

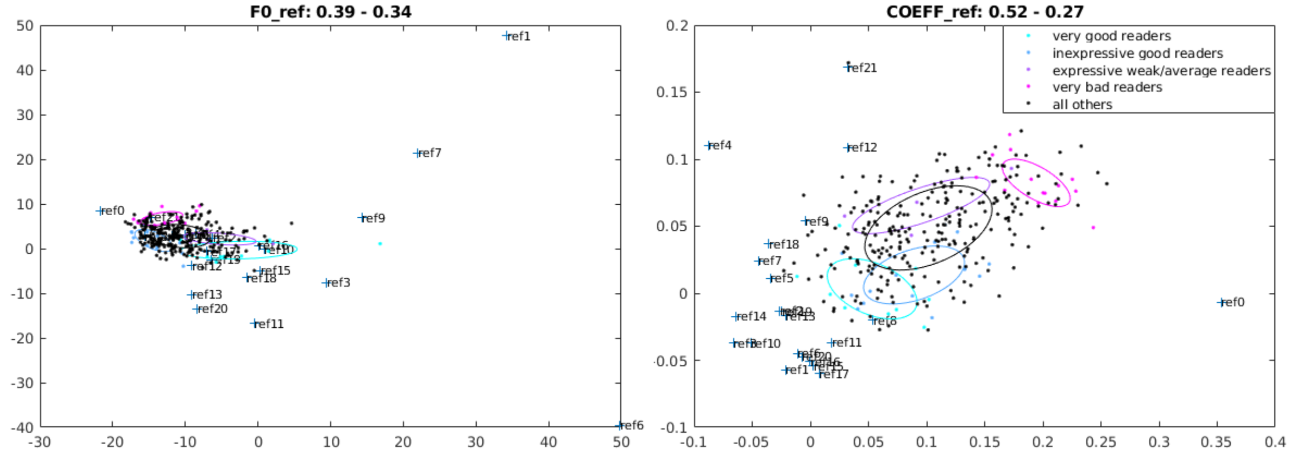


Figure 3: Projecting children' prosodic patterns onto the first MDS factorial planes (left: melody; right: rhythm). MDS is performed on 20 adult readings (points labelled as ref*). An additional landmark (ref0) with flat melody and monotonous rhythm is added for reference. Note that rhythm of very bad readers (pink ellipsis) is close to ref0 while that of very good readers get closer to adult readings.

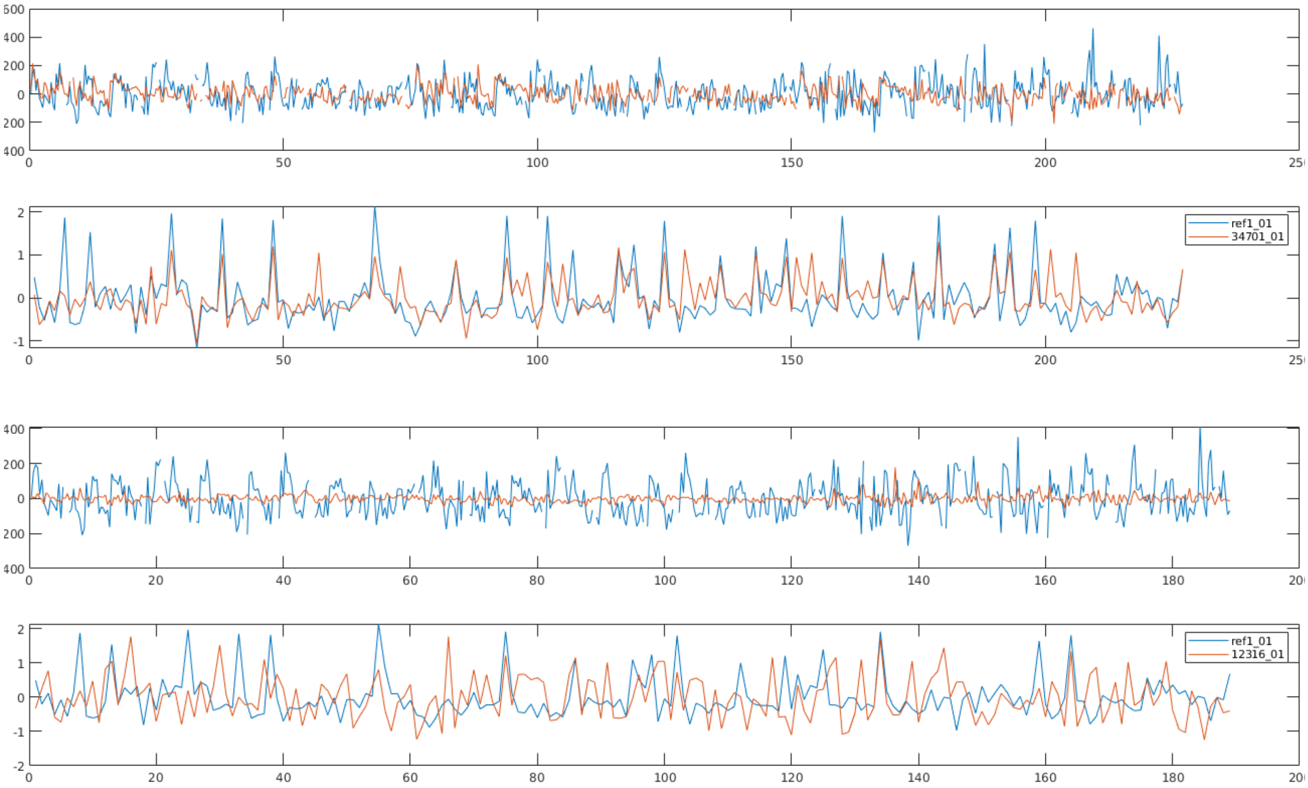


Figure 4: Alignment of centered F0 and COEFF (syllabic stretching) of a good (two top tracings) vs. bad (two bottom tracings) reader (red curves) against an adult reading (blue curves). Bad readers are characterized by flat F0 and poor alignment of pauses (they often get out of breath) while good readers better align with adult contours.

PAC (see bottom of Figure 5) while all mean absolute errors are close to .3, i.e. respectively .35, .30, .26 and .31. We further performed model simplification by iteratively removing predictors whose $\tilde{\chi}^2$ value is less than 0.05. We end up with the following formula (predictors are given with decreasing significance):

- PAC$\sim$1+CWPM+COEFF$_1$
- SMT$\sim$1+VPM+IWPM+COEFF$_1$
- PHR$\sim$1+COEFF$_1$+VPM+F0$_3$+COEFF$_3$+COEFF$_2$
- EXP$\sim$1+F0$_1$+COEFF$_1$+VPM+COEFF$_2$

Note that the main objective predictors nicely fit to their respective subjective dimensions, in particular the proposed first MDS loading factors COEFF$_1$ and F0$_1$ with Phrasing and Expressivity! Note also that COEFF$_1$ significantly contributes to the prediction of all subjective dimensions.

5. Comments

Distributions of ratings of readers with poor and excellent PAC are superimposed in Figure 5: readers with poor PAC

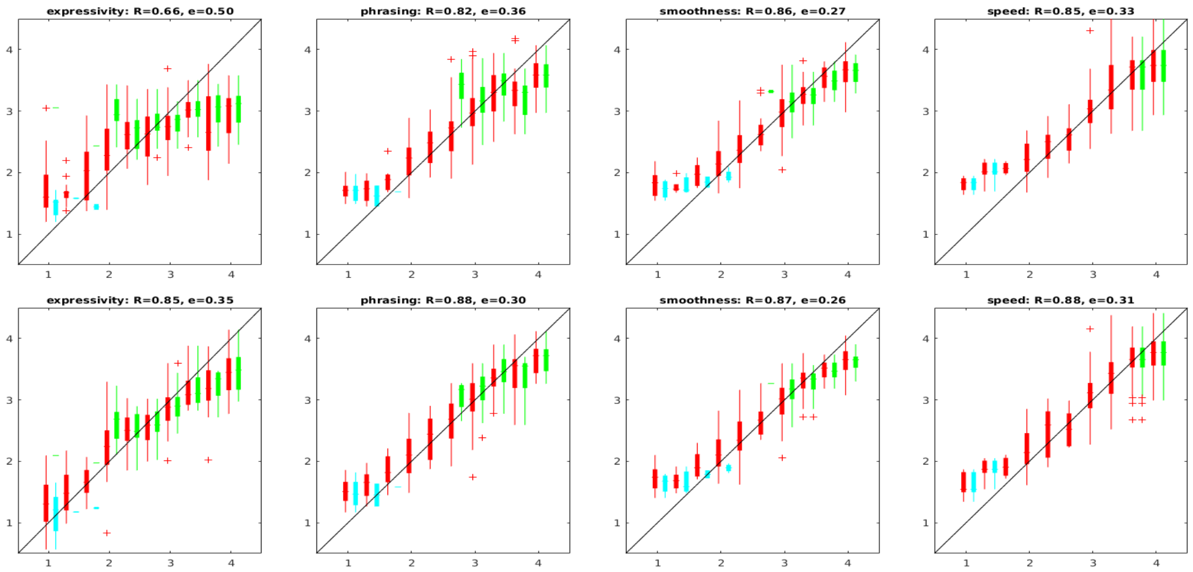


Figure 5: Predicting subjective ratings – from left to right: EXP, PHR, SMT and PAC – with no prosodic features (top) vs. with prosodic features (bottom). Distributions of ratings of readers with poor and excellent PAC are respectively superimposed in cyan and green.

are poor in all dimensions while variances of the distributions of good PAC readers increase as SMT, PHR and EXP are considered. As put forward by the review performed by Godde et al (Godde et al., 2019), the mastering of automaticity (i.e. resulting in good pace and smoothness) is a perquisite for the development of reading fluency. Expressivity is strongly linked with phrasing and comprehension, and thus opens up once syntactic and semantic processing of content can be performed online and effortless.

Note that the prosodic space calculated by MDS is yet text-specific, i.e. is expected to depend on lexical frequencies, lengths of sentences, syntactic constructs . . . experienced by readers in the chosen text. We are now collecting readings of texts with increasing complexity. The analysis of factors relating content complexity with the dimensions of the prosodic space is expected to give insights in particular aspects of reading prosody.

Finally all objective features have been hand-checked. Reproducing this performance automatically from raw signals is a challenging issue, in particular because of speech disfluencies and F0 detection of children voices.

6. Conclusions

We propose here an original technique for exploiting prosodic features in predictive models of multidimensional subjective assessment of reading fluency: melodic and rhythmic patterns uttered by young readers are compared to those of multiple reference readings performed by adults. Multidimensional scaling (MDS) is used to characterize the inter-pattern distances by few loading factors. We show that these new objective cues significantly contribute to compact and accurate predictive models.

This framework opens the pathway to the automatic assessment of reading fluency. Of course this experiment should be renewed using different text materials and fluency levels. We now perform a longitudinal study that involve an annual screening of the same pupils, in particular to assess the benefit of computer-assisted training of reading using Karaoke (Gerbier et al., 2015; Godde et al., 2017).

Acknowledgements

This work is supported by the e-FRAN project Fluence sponsored by the CDC. We thank the school teachers and pupils for welcoming Erika Godde and Manon Metz and wearing respiratory belts. Anne-Claire Dugu recorded the adult readers. Margaux Manka patiently corrected numerous phonetic alignments.

Bibliographical References

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

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

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., and Schwanenflugel, P. J. (2013). Human and automated assessment of oral reading fluency. *Journal of educational psychology*, 105(4):1142.

Breznitz, Z. (2006). *Fluency in reading: Synchronization of processes*. Routledge, New York.

Cowie, R., Douglas-Cowie, E., and Wichmann, A. (2002). Prosodic Characteristics of Skilled Reading: Fluency and Expressiveness in 810-year-old Readers. *Language and Speech*, 45(1):47–82.

Danne, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., and Oranje, A. (2005). 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*.

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

Gerbier, E., Bailly, G., and Bosse, M.-L. (2015). Using karaoke to enhance reading while listening: impact on word memorization and eye movements. In *Speech and Language Technology for Education (SLaTE)*, pages 59–64.

Godde, E., Bailly, G., Escudero, D., Bosse, M.-L., and 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)*, pages 23–27, Glasgow, Scotland.

Godde, E., Bosse, M.-L., and Bailly, G. (2019). A review of reading prosody acquisition and development. *Reading and Writing*, pages 1–28.

Godde, E., Bosse, M.-L., and Bailly, G. (submitted). Echelle multi-dimensionnelle de fluence: nouvel outil d'valuation de la fluence en lecture prenant en compte la prosodie, talonn du ce1 la 5me. *L'année psychologique*, pages 1–28.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Hirst, D., Rilliard, A., and 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*.

Launay, L. (2018). Evaleo 6-15 : batterie d'évaluation du langage oral et écrit chez les sujets de 6 15 ans. *Rééducation orthophonique*, 55(273).

Lefavrais, P. (1967). Test de lalouette.

Miller, J. and 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. *Reading research quarterly*, 43(4):336–354.

Zutell, J. and Rasinski, T. V. (1991). Training teachers to attend to their students oral reading fluency. *Theory Into Practice*, 30(3):211–217.