Children's Reading Aloud Performance: a Database and Automatic Detection of Disfluencies

Jorge Proença [1,2], *Dirce Celorico* [1], *Sara Candeias* [3], *Carla Lopes* [1,4], *Fernando Perdigão* [1,2]

[1] Instituto de Telecomunicações, Portugal
[2] Department of Electrical and Computer Engineering, University of Coimbra, Portugal
[3] Microsoft Language Development Centre, Lisbon, Portugal
[4] Polytechnic Institute of Leiria, Leiria, Portugal

{jproenca,dircelorico,calopes,fp}@co.it.pt; t-sacand@microsoft.com

Abstract

The automatic evaluation of children's reading performance by detecting and analyzing errors and disfluencies in speech is an important tool to build automatic reading tutors and to complement the current method of manual evaluations of overall reading ability in schools. A large amount of speech from children reading aloud plentiful in errors and disfluencies is needed to train acoustic, disfluency and pronunciation models for an automatic reading assessment system. This paper describes the acquisition and analysis of a read-aloud speech database of European Portuguese from children aged 6-10 from the first to fourth school grades. Towards the goal of detecting all reading errors and disfluencies, we apply a decoding process to the utterances using flexible word level lattices that allow syllable based false starts and repetitions of two or more word sequences. The proposed method proved promising in detecting corrections and repetitions in sentences, and provides an improved alignment of the data, helpful for future annotation tasks. The analysis of the database also shows agreement to government defined curricular goals for reading.

Index Terms: child speech, spoken language resources, reading performance, disfluency detection

1. Introduction

Oral reading fluency is defined as the ability to read text quickly, accurately and with proper expression [1] and it does not explicitly measure comprehension but there is evidence that oral reading fluency is an indicator of overall reading competence [2]. For the last two decades the subject of automatic literacy assessment of children has gained a significant importance and research on children's speech, its characterization, automatic recognition and evaluation has been prolific. Teachers have to consume a lot of time in the task of assessing a child's reading ability and technology can provide significant help in assisting teachers and tutors in this task. Many projects also aimed to create an automatic reading tutor that follow and analyze a child's reading such as LISTEN [3], Tball [4,], FLORA [5] and SPACE [6].

Since the speech of young children has different acoustic characteristics from adults (such as fundamental frequency, formant frequency variability, vowel duration variability, etc) [7], applying automatic speech recognition (ASR) with models trained with adult speech may be low performing and great care needs be taken to adapt or create models that target children [8, 9]. We find that there is a need to acquire a large new corpus that both satisfies the development of new recognition models for European Portuguese (EP) children's speech, and also targets the collection of the common disfluencies that children commit while reading, pursuing reading ability evaluation. Existing databases for EP such as Speecon with rich sentences [10], [11] with picture naming, the CNG Corpus targeting interactive games [12], and [13] with child-adult interaction, do not present the necessary disfluent reading speech. Therefore, we've started to collect a database of 6-10 year old children reading sentences and pseudowords, which we describe here. We have also considered that the Portuguese government defines certain curricular goals per grade, with objectives related to word and text reading, which we aim to be able to evaluate automatically.

One of the objectives of this work is to automatically detect any reading disfluencies that characterize a child's reading performance. There are several known methods for disfluency detection, such as based on Hidden Markov Models, Maximum Entropy, Conditional Random Fields [14] and Classification and Regression Trees [15], though most efforts focus on spontaneous speech. Reading disfluencies have different nuances, and certain works have targeted their automatic detection in children's reading, using complex lattice search modules or specialized grammar structures at the phonetic level [4, 16, 17] or word level context free grammars [18], though most tackle individual word reading. There is also a widespread concern in providing an overall reading ability index that is well correlated with the opinion of expert evaluators [16, 19].

For this paper we describe the acquisition of the EP children's reading aloud corpus and carry out an analysis of annotated data regarding disfluencies and reading speed metrics. Then, an attempt of automatic detection of disfluent events is defined, using word and sentence level lattices.

2. Corpus

In order to study the speech of young children reading aloud we carry out recording sessions in private and public schools. We target children that attend the primary school (1st cycle), which corresponds to an age group from 6 to 10 years old. At the time of writing, our database consists of recorded speech from 146 children. A set of 80 children's speech utterances has been fully and manually annotated in detail, including all kind of disfluencies. These 80 children, 39 boys and 41 girls, are distributed along the 4 grades as 26 from the 1st, 19 from the 2nd, 16 from the 3rd and 19 from the 4th grade. Their utterances amount to approximately 4 hours of speech. This corpus will

be increased in a near future since the annotation procedure is in progress and new acquisitions are already scheduled for the end of the school year.

2.1. Acquisition

For the acquisition process we developed an application in which the sentences are displayed in a computer screen simultaneously with the start of recording. A lapel Lavalier microphone (Shure WL93) was used as the main recording device. The recordings were performed in empty school classrooms, where background noise was minimized but cannot always be totally controlled. Young children are asked to read aloud a set of 20 sentences and 2 sets of 5 pseudowords. The pseudowords represent non-existing, non-sense words, that can be used to evaluate morphological and pronunciation awareness. The sentences vary in difficulty, accordingly to the grade years and were taken mostly from appropriate children's tales and books. The difficulty of the sentences was evaluated in terms of phonetic complexity while the pseudowords were produced by joining syllables that result in legal words in terms of pronunciation. In order to balance diversity and repetition along the database we specified 10 lists of sentences and pseudowords per grade. The vocabulary of the tasks comprises a total of 2720 words.

2.2. Disfluencies and special events

Since we are interested in the detection of reading patterns we need a database annotated in terms of disfluencies, reflecting the most common types of errors in reading aloud of children from the 1st to 4th grades.

Based on previous work [20], the annotation procedure was defined and a specific tag for each type of disfluency was assigned:

- PRE – In cases of pre-corrections, when there is a false start or total mispronunciation of the word, followed by the attempted correction.
- SUB – The word is substituted by another or suffers severe mispronunciation.
- PHO – The word experiences a small mispronunciation, usually extensions or a change in a phoneme.
- REP – The word was repeated.
- INS – An extra word that is not part of the original sentence.
- DEL – The word was not pronounced (deleted).
- CUT – The word was cut, usually in the initial or final syllable, but not corrected later.
- EXT – Extension of a phoneme.
- IWP – Intra-word pause, when a word may be pronounced syllable by syllable and silence occurs in between.

The extension and intra word pause events can occur simultaneously with others disfluencies. Other events like pauses (silence) and noises such as breathing, labial and background noise were also annotated. The number of occurrences for each type of disfluency and their percentage of total uttered words in the database are indicated in Table 1. Since the characteristics of the pseudowords are quite different from the words in sentences we opted to present a separate analysis for them.

Table 1. *Distribution of disfluency types in sentences and pseudowords (number of events and % of total uttered words.*

Tags	Sentences	Pseudowords
PRE	808 (4.77%)	136 (17.00%)
SUB	572 (3.37%)	130 (16.25%)
PHO	499 (2.94%)	168 (21.00%)
REP	381 (2.25%)	1 (0.12%)
INS	159 (0.94%)	14 (1.75%)
DEL	65 (0.38%)	3 (0.37%)
CUT	52 (0.29%)	2 (0.25%)
EXT	302 (1.78%)	95 (11.87%)
IWP	393 (2.32%)	145 (18.12%)

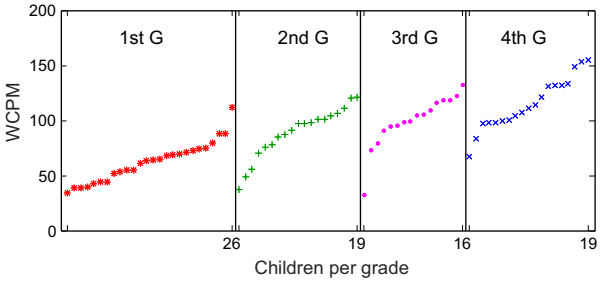

Figure 1: *Sorted words correct per minute (WCPM) per children per grade.*

3. Database analysis

From the different types of disfluencies and their frequency of occurrence in the database, we aim to determine some reading patterns for children reading aloud. The most common disfluency corresponds to a hesitation in a word followed by correction (PRE), so the child is aware of the mistake and tries to correct it. However we also find that this type of disfluency can sometimes lead to more difficult patterns when the word is subject to multiple attempts of rectification or in cases where two or more words are corrected in sequence. We can also say that children do not use filled pauses in this type of discourse, probably because this is not spontaneous speech and children tend to use a silent pause instead of a filled pause.

The words correct per minute (WCPM) parameter, which is the most widely used measure for oral reading proficiency [21], was calculated for every child. Figure 1 illustrates this parameter for the sentence reading task per grade, with sorted WCPM for each grade. To establish a comparison with the curricular goals (CG) defined by the Portuguese government, the mean value for each grade was calculated for the sentence and pseudoword tasks. The corresponding values are presented in Table 2. It can be seen that the obtained mean WCPM values for sentences are approximately coherent with the target curricular goals, which shows that the task could be appropriate for this evaluation. However, as the grade increases, the values are increasingly below the CG. This could be due to an increasing difficulty of displayed sentences along the grades, which should be adjusted. The Portuguese curricular goals for the 3rd and 4th grades do not include pseudowords; however we can see that the values for the 1st and 2nd grade are below the defined ones. In fact the 4th grade is in line with the defined CG for the 2nd grade. This variance can be justified by the higher difficulty of our pseudowords.

Table 2. *Mean words correct per minute (WCPM), SRate (the mean rate of the number of sentence words divided by the mean utterance time in minutes) and curricular goals (CG) per grade.*

	Sentences			Pseudowords		
Grade	WCPM	SRate	CG	WCPM	SRate	CG
1st	59.7	62.2	55	18.8	23.9	25
2nd	85.2	88.8	90	26.7	31.1	35
3rd	97.1	99.7	110	26.1	30.5	-
4th	104.1	115.1	125	34.9	38.0	-

To conclude this analysis, we calculate the percentage of words that present any related disfluencies, which are, from the first to fourth grade, 18.67%, 13.27%, 10.14% and 11.04%. This means that, in average, for the first grade, approximately for every 5 words, one of them presents at least one disfluency of any type.

4. Automatic detection of disfluency events

The two main disfluency events that provide extra segments in the transcribed utterances are wrongly or partially pronounced words followed by correction (represented by the PRE tags) and repetitions (REP), with the other being the less represented insertions. We propose a method that aims to not only detect these two events, but also to provide a good time-alignment of the database in terms of existing segments. This means that, for now, we are not considering pronunciation errors, which would require detailed pronunciation models that we aim to develop. In terms of alignment, all the SUB and PHO tags are replaced by their corresponding word, hoping that the phonetic decoder will still match the word to the mispronounced segment.

We parameterize the dataset in standard Mel-Frequency Cepstral Coefficients (MFCC), and apply a voice activity detection to counter intra-word pauses. The detection of segments is performed through Viterbi decoding of the utterances using specifically built lattices and Hidden Markov Models (HMMs) as the acoustic model. The phonetic models used were trained with 3 hours of the annotated utterances, and are standard triphone HMMs with 10500 Gaussians. Further aspects of the voice activity detection step and the task lattices are specified below.

4.1. Voice Activity Detection

There are children, especially first graders, who often pronounce words syllable by syllable. This may lead to silence intervals between syllables, problematic for the automatic word decoder which could mark several events with silence in between instead of only a full word. With the fairly low-noise environment of our dataset, our approach to this problem was to find and cut any segments of low-energy signal and perform decoding with a cut utterance. We decided to simply analyze the large-window logarithmic Energy of a signal and consider a threshold below which a candidate silence segment must fall. Only segments with a certain minimum duration (220ms) are then cut from the original signal and their start and end times saved. Signal cutting is performed at MFCC frame level, so the new transitions created are not too problematic.

Existing silence does help to more clearly separate some words for the decoder but we found this approach to be helpful

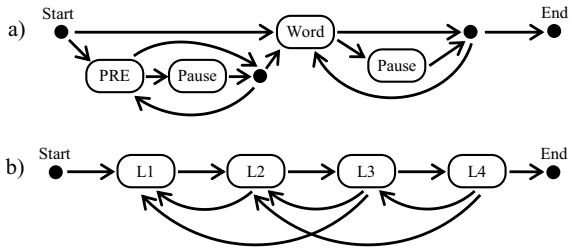

Figure 2: *Schematic of an individual word sub-lattice (a) and example of a final lattice for a 4 word sentence using 4 word sub-lattices L1 to L4 (b).*

in the dataset. At the end of decoding, the obtained markers for the cut utterance are expanded to the original duration of the signal, by reconstructing the low-energy segments. If one of these segments falls deep within a tagged word, the word is marked has having silence between syllables. If a segment falls very close (100ms) within the beginning or end of a decoded word, it is probable that the small mismatched portion belongs to an adjacent word (note that these words should be originally separated by a significant pause). Thus, the first word is compacted to start or end at the segment's time, and the adjacent word expanded.

4.2. Task Lattices

To detect the extra events related to a word, a special word lattice is built as input for the decoder. Each sentence will have a specific lattice, and the starting point is the full original sequence of words as shown to the reader. From this, each word is represented by a sub-lattice where PRE and REP events are possible as described in Figure 2a. We decided to code PRE events as all sequences of syllables that a word can be divided as, always starting at the first one and up to the second to last. For example, for a 4 syllable word, the possible PRE events are: syl1; syl1+syl2; syl1+syl2+syl3. These possibilities are comparable to most PRE cases where a word is pronounced (correctly or not) up to an interruption point, followed by a subsequent correction. Pause events (for silence, respiratory events or noise) are allowed between all word-related events and are optional. Sequences of these sub-lattices would not account for repetitions or corrections of sequence of words, so the final sentence lattice is defined to allow a back step of one or two word sub-lattices as pictured in Figure 2b.

The probabilities/weights of the arcs that allow PRE events, repetitions of PRE and repetitions of words were defined as very low values (minimum of 1%). These probabilities could be derived from the data in some way, but all trials so far (word position in the sentence, word difficulty) provided practically similar results.

4.3. Results

The baseline system used for comparison is simply a forced alignment of the exact sequence of words of the original sentence presented to the child, with optional paused segments between words. Comparing the word-level alignment in terms of the word sequence to the manually tagged reference, the baseline is essentially at fault due to non-detections of extra events (deletions). Results are presented in Table 3 where the Final System is the one that presents 2.6% false alarm rate, and it can be seen that the baseline already surpasses 90% accuracy. In fact, around 60% of sentences do not present any PRE or REP event, and are considered fully correct in this

Table 3. Word sequence alignment Correctness and Accuracy. Percentage of segments that match in label and in boundaries with several tolerance collars.

	Baseline	Final System
Word Corr. %	92,18	97,83
Accuracy %	91,86	94,87
50 ms collar %	55,72	57,86
100 ms collar %	70,05	82,31
150 ms collar %	76,05	89,21
200 ms collar %	79,99	92,12
250 ms collar %	82,76	93,63

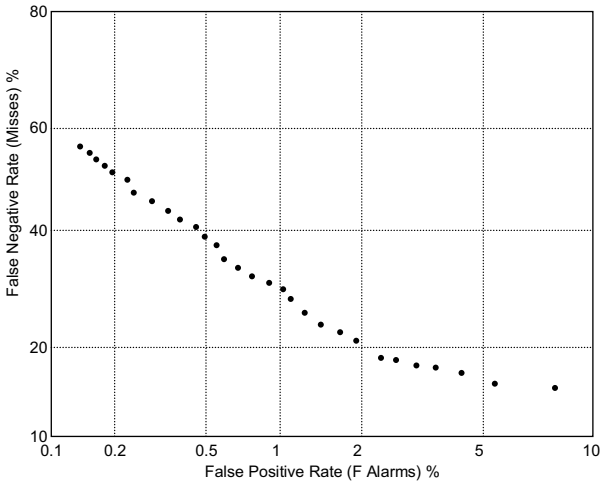

Figure 3: *Detection Error Tradeoff (DET) curve for the disfluency detection system with varying word insertion log probabilities.*

analysis. To analyze the performance of our system in detecting 1326 PRE and REP events, we consider that: any extra detected word events that are not aligned with reference are false alarms; failing to detect the occurrence of a PRE or REP segment is a miss; erroneously detecting a PRE or REP segment as an event of an adjacent word is also a miss (e.g., a PRE event in reference was detected as a repetition of the previous word). With this, we iterate the system using different word insertion penalties at the decoding stage, leading to changes in miss and false alarm rates presented as a Detection Error Tradeoff (DET) curve in figure 3. For a 5% false alarm rate we obtain 15% miss rate (corresponding to +5 word insertion log probability). Since there is often phonetic mismatch on the semi-forced alignment due to mispronunciations, the system can mix repetitions of one word with pre-events of the next one, which we counted as a miss. Thus, hypothetically considering these substitutions as detections (since they are also hesitation events), the miss rate would fall to around 6.9%. Furthermore, there are 173 insertion tags that are never accounted for, lowering overall accuracy.

To verify if the time boundaries of the automatically decoded words and events are similar to the manual reference, we analyze the percentage of correctly detected segments, given that each boundary falls within a tolerance of increasing collar thresholds (Table 3). Only with large collars it is noticeable the improved match from the output of the proposed system. Even with a collar of 250 ms we do not reach the obtained correctness when only the transcription sequence is considered. There are several factors that can explain the discrepancy in alignment with the manual reference. We found that the pause markers on the manually annotated reference can be very strict and are very often mismatched to system output, influencing words and events before or after a pause. Also, improved acoustic models for European Portuguese children would be greatly helpful, as the amount of data used to train the used models could be low.

Since we didn't target or account for correctness of pronunciation, this system should be seen as a proof of concept of detecting extra events through what can be called as a semi-supervised decoding, or a forced alignment with certain freedoms. Other aspects of our system can account for errors in specific event labeling: for small words with 1 syllable only repetitions are marked, and since some PRE tags of larger words are mispronunciations of the whole word, they can be decoded as the word followed by repetitions.

5. Conclusions

A European Portuguese database of children's reading was collected, with the objective of analyzing common error and disfluency events in reading tasks performed by children. Many types of disfluencies were identified and we notice the lack of some hesitations characteristic of spontaneous speech such as filled pauses. Although we increased the difficulty of sentences along the grades, the average metric of correct words per minute fell close to the government defined curricular goals, showing the suitability of the task for this evaluation. The pseudowords reading task was probably not adequate for the intended goals, and its difficulty may have to be reviewed.

A system was developed for automatic detection of extra events of repetitions and corrections in sentences. It considers the case where sequences of words are repeated, such as a child starting a sentence from the beginning if a mispronunciation happens on the second or third word. The results show promise in detecting false starts and repetitions (e.g., 15% miss and 5% false alarm rate), and although the system should undergo several improvements, it certainly and immediately fulfils one of our initial goals: to provide an improved aligned transcription as a starting point for the arduous task of manually tagging and aligning all events in future extensions to the dataset. Also, the task of sentence reading leads to less predictable disfluencies, compared to, e.g., isolated word reading. The proposed sub-lattice based approached could be suffering from not using more robust acoustic models of European Portuguese children, which we are refining as the dataset increases. We intend to evaluate alternative methods, such as approaching the problem as a word spotting task. The final goal of this work is to detect mispronunciations and other disfluencies in a child's set of reading tasks and provide an overall reading ability index that correlates well to the evaluation of human experts.

6. Acknowledgements

This work was supported in part by Fundação para a Ciência e Tecnologia under the projects UID/EEA/50008/2013 (pluriannual funding in the scope of the LETSREAD project) and CATARATA PTDC/DTP-PIC/0419/2012; and Marie Curie Action IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP). Jorge Proença is supported by the SFRH/BD/97204/2013 FCT Grant. We would like to thank the *João de Deus* and *Bissaya Barreto* school associations and CASPAE parent's association for collaborating in the database collection.

7. References

[1] National Reading Panel, "Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction." National Institute of Child Health and Human Development, Tech. Rep. 00-4769, Washington, DC, 2000.

[2] L. Fuchs, D. Fuchs, M. Hosp, and J. Jenkins, "Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis." *Scientific Studies of Reading* 5, 239–256, 2001.

[3] J. Mostow, S. Roth, A. G. Hauptmann, and M. Kane, "A Prototype Reading Coach that Listens," *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), American Association for Artificial Intelligence, Seattle, WA*, August 1994, pp. 785-792.

[4] M. Black, J. Tepperman, S. Lee, P. Price, and S. Narayanan, "Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment," *Proc. InterSpeech, Antwerp, Belgium*, 2007, pp. 206-209.

[5] D. Bolaños, R. A. Cole, W. Ward, E. Borts, and E. Svirsky, "FLORA: Fluent oral reading assessment of children's speech," *ACM Transactions on Speech and Language Processing (TSLP)*, 7(4), 16, 2011.

[6] J. Duchateau, Y. O. Kong, L. Cleuren, L. Latacz, J. Roelens, A. Samir, K. Demuynck, P. Ghesquière, W. Verhelst and H. Van hamme, "Developing a Reading Tutor: Design and Evaluation of Dedicated Speech Recognition and Synthesis Modules," *Speech Communication*, vol. 51, no. 10, pp. 985-994, October 2009.

[7] S. Lee, A. Potamianos and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameter," *Journal of the. Acoustical Society. of América* 105, 1455, 1999. http://dx.doi.org/10.1121/1.426686

[8] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, Nov. 2003.

[9] A. Hämäläinen, S. Candeias, H. Cho, H. Meinedo, A. Abad, T. Pellegrini, M. Tjalve, I. Trancoso and M. S. Dias, "Correlating ASR Errors with Developmental Changes in Speech Production: A Study of 3-10-Year-Old European Portuguese Children's Speech," *In WOCCI 2014 – Workshop on Child Computer Interaction, Singapore*, September 2014, pp. 7-11.

[10] The Speecon Portuguese Database http://catalog.elra.info/product_info.php?products_id=798

[11] C. Lopes, A. Veiga and F. Perdigão, "A European Portuguese Children Speech Database for Computer Aided Speech Therapy," *In Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012, LNCS,* vol. 7243, pp. 368–374. Springer, Heidelberg, 2012.

[12] A. Hämäläinen, S. Rodrigues, A. Júdice, S. M. Silva, A. Calado, F. M. Pinto and M. S. Dias, "The CNG Corpus of European Portuguese Children's Speech," In *Text, Speech, and Dialogue*, pp. 544-551, Springer Berlin Heidelberg, January, 2013.

[13] A. L. Santos, M. Génereux, A. Cardoso, C. Agostinho and S. Abalada, "A corpus of European Portuguese child and child-directed speech," in *Proceedings of the 9th Conference on Language Resources and Evaluation – LREC 2014, European Language Resources Association (ELRA),* 2014.

[14] Y. Liu, E. Shriberg, A. Stolcke and M. Harper, "Comparing HMM, maximum entropy, and conditional random fields for disfluency detection," in *Proc. Interspeech*, 2005, pp. 3313–3316.

[15] H. R. B. Medeiros, H. Moniz, F. Batista, L. Nunes and I. Trancoso, "Disfluency Detection Based on Prosodic Features for University Lectures," in *Interspeech 2013, ISCA, Lyon, France*, August 2013, pp. 2629-2633.

[16] J. Duchateau, L. Cleuren, H. Van Hamme, and P. Ghesquière, "Automatic assessment of children's reading level," in *Proc. Interspeech, Antwerp, Belgium*, Aug. 2007, pp. 1210-1213.

[17] E. Yilmaz, J. Pelemans and H. Van hamme, "Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model," *INTERSPEECH*, 2014, pp. 969-972.

[18] X. Li, Y. C. Ju, L. Deng and A. Acero, "Efficient and robust language modeling in an automatic children's reading tutor system," in *Acoustics, Speech and Signal Processing*, *ICASSP 2007. IEEE International Conference*, April 2007, vol. 4, pp. 193-196.

[19] M. P. Black, J. Tepperman and S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *Audio, Speech, and Language Processing, IEEE Transactions*, 19(4), 1015-1028, 2011.

[20] S. Candeias, D. Celorico, J. Proença, A. Veiga and F. Perdigão, "HESITA(tions) in Portuguese: a database," in *DiSS 2013, ISCA endorsed Interspeech 2013 satellite workshop*, *August 21-23, KTH Royal Institute of Technology, Stockholm, Sweden*, 2013, pp. 13-16.

[21] J. Hasbrouck and G. A. Tindal, "Oral reading fluency norms: A valuable assessment tool for reading teachers," *The Reading Teacher*, 59(7), 636-644, 2006.