LETSREAD

Hugo José de Sousa Ferreira

System for Automatic Assessment of Reading Ability of Children

September 2016

• U C •

Universidade de Coimbra

Department of Electrical and Computer Engineering

Faculty of Sciences and Technology, University of Coimbra

Dissertation for Master degree in

Electrical and Computer Engineering, field of Telecommunications

System for Automatic Assessment of Reading Ability of Children

Hugo José de Sousa Ferreira

Supervision by:

Prof. Fernando Manuel dos Santos Perdigão, PhD

Co-Supervision by:

Jorge Daniel Leonardo Proença

Jury

President: Prof. Luís Alberto da Silva Cruz, PhD

Member: Prof. Carla Alexandra Calado Lopes, PhD

Member: Prof. Fernando Manuel dos Santos Perdigão, PhD

September 2016

iii. Acknowledgments

Since this is a personal note the acknowledgements will be written in the mother tong of the author, Portuguese.

Agradeço de uma forma muito especial ao Prof. Doutor Fernando Perdigão pela paciência, motivação e disponibilidade empreendida nesta dissertação. Foi um privilegio desenvolver esta dissertação sobre a sua orientação.

Agradeço a todos os membros do laboratório de Processamento de sinal, em particular ao Jorge Proença pela grande ajuda e conselhos transmitidos no decorrer desta dissertação.

Aos meus amigos que comigo percorram este longo percurso académico. Aos amigos mais distantes e aqueles que mesmo espalhados pelo mundo ajudaram-me sem o aperceberem.

Por ultimo e de maior importância agradeço à minha família, especialmente ao meu pai, mãe e irmã que sempre me apoiaram e me deram o sentimento mais importante, amor.

A todos um profundo agradecimento.

iv. Abstract

Reading fluently is a complex activity and challenging learning process, that requires multitasking and an almost immediate speech response. Proper reading fluency is a crucial skill for a human being in a modern society, and it is mostly acquired in primary school.

Reading fluently is characterized by an accelerated, seemingly effortless, autonomous and unconscious process. Oral reading fluency is manly express in three different components, reading rate or speed, reading accuracy and reading expressiveness.

For a child between 6 and 10 years old to achieve the curricular goals, precious teacher's working time must be dispended with each child. In this point science and technology comes to the rescue. A system, used as a tutor, training or evaluator, to automatically assess the oral reading fluency of children's can save the teachers precious time and at the same time improving the children's reading ability.

Children tend to incur in several disfluencies, which are normal since they are still learning to read. A system to automatically assess a child's reading ability must be able to distinguish correctly pronounced words from those with disfluencies, and also to provide with features that proper characterize oral reading fluency.

Keywords: Features, Children's Reading Ability, Posteriorgrams, Detection of Disfluencies, Phonemes Recognizers.

v. Resumo

Uma leitura fluente é uma atividade complexa e processo de aprendizagem difícil, que requer o recurso a múltiplas tarefas e necessita de uma resposta quase imediata. A fluência de leitura adequada é uma habilidade crucial para um ser humano na nossa sociedade moderna, e é principalmente adquirida na escola primária.

Uma leitura fluente é um processo caracterizado por um ritmo acelerado, aparentemente sem esforço, autônomo e inconsciente. A fluência oral de leitura é expressa e caracterizada por três componentes diferentes, o ritmo de leitura ou a velocidade da mesma, a precisão na leitura e sua expressividade.

Para uma criança entre 6 e 10 anos alcançar as metas curriculares proposta é necessária uma tremenda perda de tempo de trabalho do professor com cada criança ao seu encargo. Neste ponto a ciência e tecnologia está pronta para o resgate. Um sistema, usado tanto como um tutor, ou avaliador, que avalie automaticamente a fluência de leitura das crianças pode salvar aos professores precioso tempo e ao mesmo tempo melhorar a capacidade de leitura das crianças.

As crianças tendem a incorrer em várias difluências, que são normais uma vez que ainda estão a aprender a ler. Um sistema para avaliação automática da capacidade de leitura de uma criança deve ser capaz de distinguir corretamente as palavras bem pronunciadas das palavras com difluências e também para fornecer estatísticas que caracterizem adequadamente fluência oral de leitura.

Keywords: Features, Capacidade de Leitura de Crianças, Posteriorgramas, Deteção de Difluências, Reconhecedores de Fonemas.

1 Contents

vi. List of Figures

vii. List of Tables

viii. List of Acronyms

ANN - Artificial Neural Network

HMM – Hidden Markov Models

DET – Detection Error Tradeoff

EER – Equal Error Rate

LLR – Log-Likelihood Ratio

MLP – Multilayer Perceptron

1 Introduction

Children start to read in primary school with a teacher's support. The use of automatic speech recognition to support in this task can be a crucial development for a child improve his reading ability.

The constant improvement of this methods can use the child's speech to detect events, like words, phonemes, pauses, etc. To evaluate child's ability to read it should be determined the oral reading fluency presented in the utterance (Fuchs et al, 2001). The oral reading fluency can be assessed by a different set of characteristics, like words correct per minute.

Is not necessary to evaluate the oral reading fluency determine if the text was understood, even if it can also be a good indicator. Henceforward oral reading fluency can be evaluated using automatic speech recognition. A good indicator of proper reading fluency is absence of disfluencies in the utterances and it that can be detected using detection methods.

There are several detection methods, one detection method is the use of a trained artificial neural network to map the probabilities of phonemes occurrences.

1.1 Motivation

This dissertation inserts itself in the Letsread project, which is a project that intends to automatically evaluate the Portuguese children's reading ability. the main problem in evaluating a child's reading ability is properly detecting disfluencies in words.

The motivation behind this dissertation is to correctly assess a child's reading ability using characteristics from a child's speech resorting automatic speech recognizers.

1.2 Objectives

This dissertation objective consists in automatically evaluate a child's reading ability. The objective is to for a given audio speech of a child's reading a sequence of words, the system should be able to determine the reading ability index, by automatically align the phonemes or words with the recorded audio. Then by resorting to a feature extraction, like correct words per minute, correctly evaluate the child's reading ability.

1.3 Dissertation Organization

This dissertation is organized in 5 chapters. The first chapter provides with an introduction to this work and describes its objectives and the motivation. The second chapter is related with the database used in this works, also with an overview of the children's reading ability. The third chapter describes the methods used to automatic recognized a speech utterance. The fourth chapter describes the steps taken until the final proposed model, with a description of the features extracted from the recognizers in order to characterizes the child's reading speech. Also describes the resulting models and provides with a few examples of the system. The fifth chapter presents with a conclusion of this dissertation.

2 Children's reading

This chapter focussess in providing an overview of the speech and reading ability of children. Also gives an overview on the Let's Read project and its respective databases used in this dissertation.

.

2.1 Children's Reading Assessment

Speech is affected by many factors, in which one of them is the age of the speaker. Children's speech differs in many ways from an adult speech, like the higher pitch or the difficult gender differentiation.

2.1.1 Speech

Speech is the vocalized form of human communication. It is a unique and staggering ability that sets humans apart from other animals and still a huge mystery in the human's evolutionary road. By a biological manipulation of the vocal tract a different set of non-stationary acoustic signals are produced within a range of frequency from 300Hz to 3400Hz. Speech is based in a syntactic combination of grammar and lexicon drawn from a vocabulary, a large set of words, that compose a language (Speech, wiki).

Any language can be dismembered into a few basic units that distinguish a word from another, the phonemes. The acoustic realization of a phoneme is a phone. Phoneme and phone, while intrinsically connected, are slightly different concepts. Phonemes are absolute to all languages and can all be represented by the International Phonetic Alphabet. The acoustic realization of the phonemes depends on many factors, such as the vocal tract of the speaker, the type of speech (colloquial, formal, etc) or even the acoustical environment (Lopes, 2011).

A phoneme it is an abstract concept that tries to catalogue distinctive unit sounds in a particular language. For example, in Portuguese the words "carro – /k a R u/" and "caro – /k a r u/" have only one different phoneme that changes the whole meaning of the word, the phoneme /R/ or /r/, see Appendix A.

Nevertheless, speech can be viewed as a chain of phonemes. The technique to obtain the phonemes from the speech signal is named phoneme recognition, which plays a fundamental role in this work.

2.1.2 Children's Reading Ability

*"Reading is a complex performance that requires simultaneous coordination across many tasks. To achieve simultaneous coordination across tasks, instantaneous execution of component skills is required. With instantaneous execution, reading fluency is achieved so that performance is speeded, seemingly effortless, autonomous, and achieved without much consciousness or awareness" (*Logan, 1997*).*

To assess a person's reading level and proficiency, Oral Reading Fluency (ORL) is a good indicator of reading competence. *"Oral reading fluency in children is defined as the ability to read text quickly, accurately and with proper expression" (*National Reading Panel, 2000*).* It is understood that from a behavior point of view Oral Reading Fluency is directly linked to rapid word recognition, basically if a person understands the word that he is reading, his reading skills will improve. However, this dissertation focuses only and exclusively in evaluate a children's reading ability without considering children's comprehension on the read text (Fuchs et al, 2001).

2.1.3 The LeastRead Project

As referred before this dissertation inserts itself in the Letsread project. The project is an internal project of the telecommunications institute of the electrical engineering department of university of Coimbra. The Letsread project intends to modules the Portuguese children's reading ability. By developing models to detect disfluencies and other occurrences in the children's speech is possible to automatically evaluate a child's reading ability.

2.2 The LetsRead Database

In the scope of Letsread project a speech database has been collected in several elementary schools of the Coimbra area. It corresponds to recordings of children reading several sentences and pseudowords. Pseudowords are words that are not part of the lexicon of the language (Portuguese) but are pronounceable and useful to measure the phonological awareness in reading.

The LetsRead database contains 7418 audio recordings of read-aloud utterances by 284 children, students, from first through fourth grade (Proença et al,2015). Each student read aloud a set of 20 sentences and 10 pseudowords. Note that some students read 5 pseudowords per recording while others read 1 pseudoword per recording. There were 4 recording collections, on July 2014 December 2014, May 2015 and June 2015, resulting in a total of 20 recording hours.

A total of 2268 recordings, from 104 students and regarding only the first two sets of recordings, were manual annotated. The annotation is very detailed indicating all disfluencies and the time intervals were each word has been uttered.

To determine a child's reading ability index a total of 1200 recordings have been used, this 1200 recordings were also manually annotated and are described below.

Farther in this dissertation both manually annotations will be used in different points as a ground truth to evaluate the used methods.

2.2.1 Manual Annotation of LetsRead Database

The collected speech of this database has a wide variety of disfluencies and reading errors, which are normal in reading aloud. The disfluencies consist in hesitations, pauses, as well as syllabifications, mispronunciations and change of some words. These events were labelled in the manual transcription in the followed way (Proença et al,2015).

For each type of disfluency, a specific tag was assign to it:

PRE – in case of a false start or total mispronunciation of the word followed by a correction attempt;

SUB – in case of word substitution;

PHO – in case of a word extension or a phoneme exchange; REP – in case of a word repetition; DEL – in case of an unpronounced word; CUT – in case of a cut word and no correction after; EXT – in case of a phoneme extension; IWP – in case of an intra-word pause, when a word is pronounced syllable by syllable and silence occurs between.

The fully annotated database has 36% of disfluency events. The disfluency PRE is the most common in the utterances, with 5.14% events of the total uttered words. In the pseudowords, although PRE presents a high number of occurrences, the most common disfluency is SUB with 22.4% events of the total uttered words, which makes sense considering the fact that those words are not familiar to the children, but they have to pronounce them correctly.

2.2.2 The Teachers Evaluation Process of LetsRead Database

In order to have a reference measure of the children's reading ability, an evaluation process with experienced teachers has been conducted in the scope of the Letsread project.

For grading purposes 150 children were selected for grading by 105 experts, 43 children from the first, 40 from the second, 35 from the third and 32 from the fourth grade. Each expert grade a set of 15 children, resulting in 7 to 13 evaluations for each child. The experts grade each child individually after listening 6 audio files, five with sentences and one with 5 pseudowords, proving the respective child with a score ranging from 0 to 5. The grading system as the followed meaning:

0 – corresponds to below first grade level;

1 – corresponds to the end of first grade level;

2 – corresponds to the end of the second grade level;

3 – corresponds to the end of the third grade level;

4 – corresponds to the end of the fourth grade level;

5 – corresponds to above the fifth grade level.

A mean of reading ability scores for each child was obtained. However due to the fact that between the scorers there is some classification unconformity, pair wise correlation and correlation to mean was used to determine the invalid scorers, using the mean value per each child as a ground truth. For feature analysis it was used the Z-normalization values, an alternative method to using a mean score for child from raw values. The z-norm per evaluator took some bias effects that can be, for instance, a teacher constantly giving higher or lower scores than the average ones; a teacher constantly giving scores near the maximum or the minimum; or constantly giving middling scores. This metric will be designated as z-norm index.

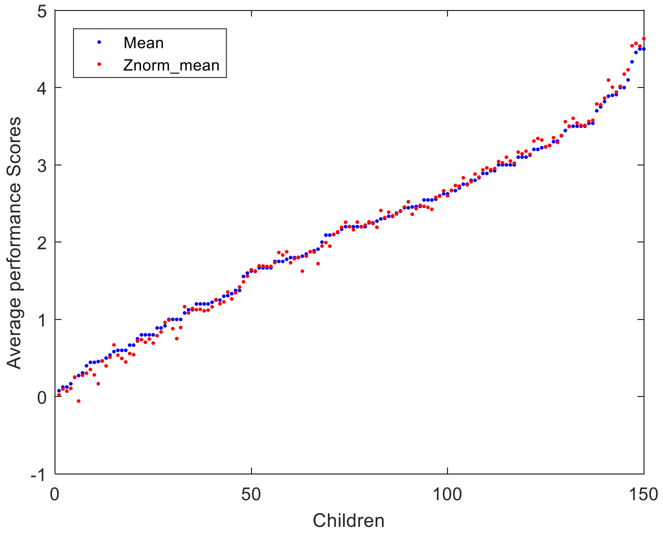

Figure 1 – Mean indexes vs z-norm indexe

In the figure 1 there is a comparison between the final scores given by the z-norm method to the mean values of raw scores. It is visible that z-normalization method can give values a bit lower than 0 or a bit higher than 5.

3 Background and methods of automatic speech recognition

This chapter summarizes some important background theory behind the obtainment of Posteriorgrams used to devolved the work presented such as Artificial Neural Network (ANN), Hidden Markov Model (HMM). Also covers the methods used for phonemes and words recognition using posteriorgrams.

3.1 Models for Speech Processing

A set of Np Phonemes, referring to the Portuguese language, will characterize the phones realization, and will be recognized by phonemes recognizers.

3.1.1 The Posteriorgram concept

A phoneme recognizer gives a matrix of probabilities à posteriori of the phonemes in each frame, which is often designated as Posteriorgram. The Posteriorgram indicates the à posteriori probabilities of a *pn* phoneme for a given *xt* frame of the signal, Pr(pn|xt). The Posteriorgram is a matrix of Np phonemes by T frames (Np*T), were each frame has à posteriori probability of each phoneme which the frame sum gives 1 with probability values from zero to one.

The Posteriorgram is generated by an Artificial Neural Network (ANN) and it is used as a base for determine a phoneme or word sequence. The ANN uses Hidden Markov Models to model each phoneme with three states, resulting that each phoneme of the Posteriorgram possesses 3 states.

Figure 2 shows an example of Posteriorgram representation with probabilities quantized in colours, where red possesses high probability and blue low probability. The frames are represented in the abcissas and the phonemes in the ordinates.

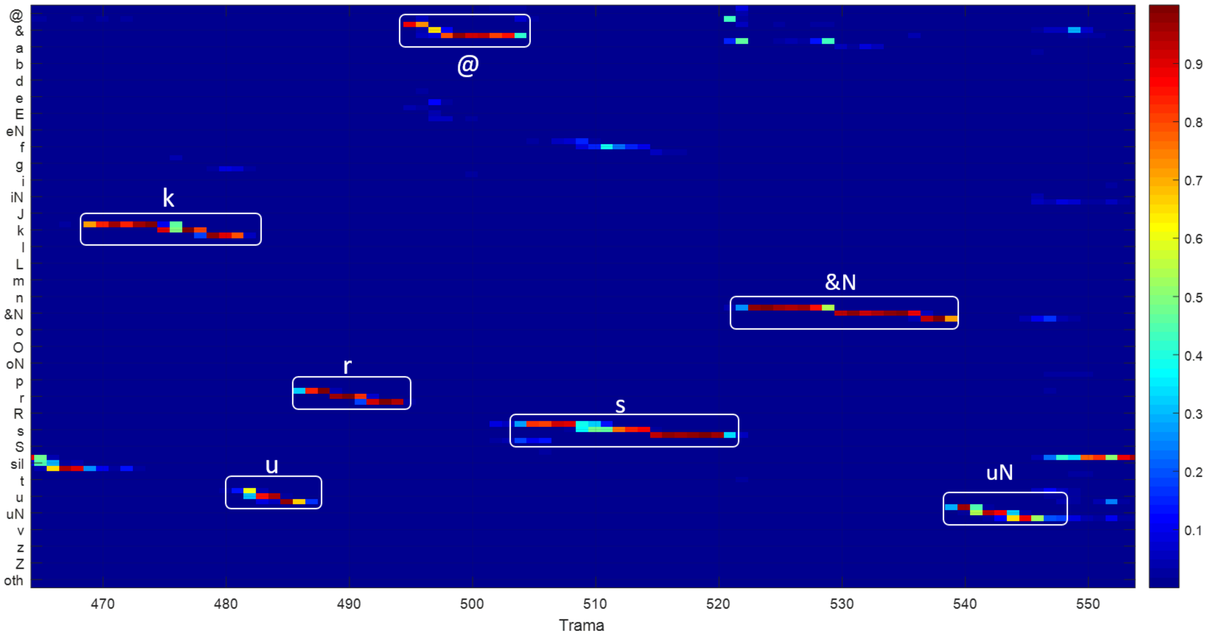

Figure 2 – Graphic representation of à posteriori probabilities matrix example

Figure 2 as a graphic representation of a posteriorgram in which the word “coração [k u r & s &N uN]” was pronounced.

3.1.2 Artificial Neural Networks

The concept of Artificial Neural Networks was inspired by Biological Neural Networks such as the ones presented in our human brain. a huge network of interconnected nodes, such as our brain, bases her knowledge in a learning process that uses the connection forces between those nodes to storage the experimental knowledge in what's called as synaptic weights. ANN, just like our brain, is a wonderful learning machine with a lot of applications in computational machine learning. In speech recognition there's a widely accepted ANN architecture, the Multilayer Perceptron- MLP (Castela,2015).

MLP is a ANN architecture characterized by being a feedforward network, hence there's no feedback from the subsequence layer to the preceding layer. The network is typically structured with an input layer formed by a set of input nodes, with a hidden part composed by one or more hidden layers formed by computational nodes and with an output layer of nodes containing the network information results (Castela,2015). Figure 3 shows an example of a MLP from (Saracoglu,2010).

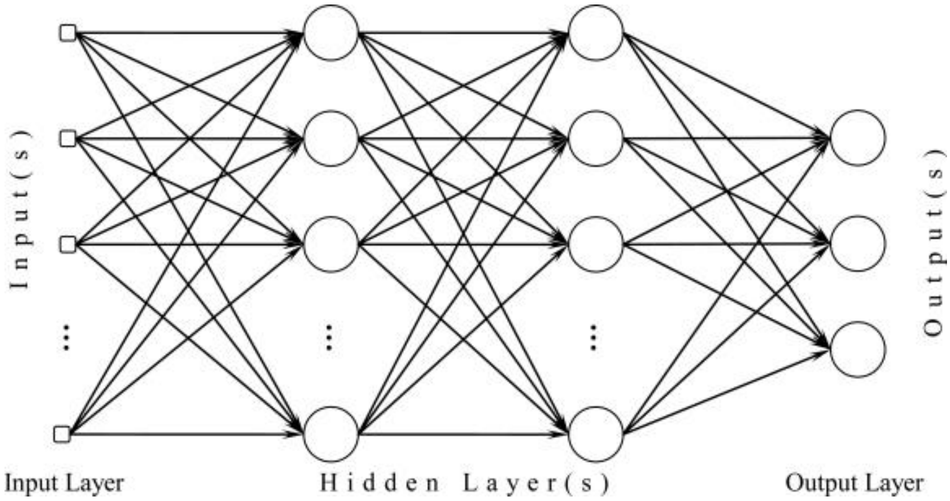

Figure 3 – Overview of MPL arcquitecture. Taken from (Saracoglu,2010)

This architecture uses as a learning technique the *error back-propagation algorithm*, that allows the network to calculate the error at the output layer. By a propagation in the conventional sense (*forward pass*), from the input layer through the hidden layers, the network maps the nodes reaction to the input information giving a response at the output layer, then an inverse propagation (backward pass) with the error information, in this case the difference between the phonemes transcription and the network obtained mapping, that will adjust the synaptic weights towards the optimal response.

The outputs of the Artificial Neural Network are 3*Np phonemes which are the states of the considered phonemes with the insertion of an additional phoneme "oth".

In this work it is used a 35 and 40 phonemes network, 102 states plus 3 states for the “oth” phoneme and 117 states plus 3 states for the “oth” phoneme respectively, developed by (castela,2015) and also a 41 phonemes network, 120 states plus 3 states for the “oth” phoneme, that was under development in the duration of this dissertation by (Franco,2016). This 41 phonemes network was still under development and the results described are not the final ones, since it is expected that this network surpass the other two.

3.1.3 Hidden Markov Models

In speech recognition HMM are widely used and currently accepted as the technique with the best performance. A speech signal is not a stationary process due to the fact that our vocal tract produces different frequencies in a time-based signal, however a speech signal can be viewed as a piecewise stationary signal or short-time stationary signal (Paul, 1990). This means that dividing a speech signal in frames of a short time-scale (about 10 milliseconds), the signal under the fame can be approximated as a stationary process and modeled with its spectral characteristics. The sequence of frames can be modeled under the framework of a Markov chain with its (hidden) states having a particular probability distribution.

Figure 4 shows an example of a HMM. An HMM is defined by a set of *N* states, a matrix of state transition probabilities, a_{ij}, and the probability distribution function of each state, $b_j(o)$. The filled circles indicate non-emitting states. In the figure, the emission of four observations (o_1, o_2, o_3, o_4) is indicated, the first two in the first emitting state.

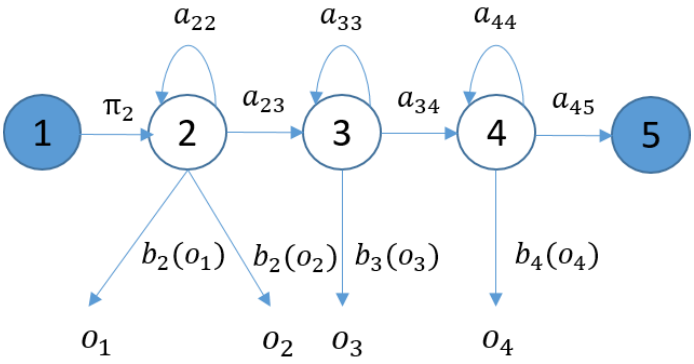

Figure 4 – HMM example for a 3 states left-to-right

There are a number of possible variations of the HMM. In speech recognition, as well as in this work, usually a 3 states left-to-right state model is used.

3.2 Decoding Methods

In this work two different methods for decoding the speech were used: as a sequence of phonemes (using a free-phone-loop or a bigram) and as a sequence of words (by aligning the speech signal with the words of the prompt - the read sentence). These methods are named henceforward as Phoneme Recognizer and Word-Alignment, respectively. Both methods are based on the Viterbi's Algorithm with the use of Token-Passing Algorithm, using the posteriorgrams generated by the neural network.

3.2.1 The Phoneme Recognizer

The first method applied for posteriorgram decoding is a phoneme recognizer. The output of this recognizer is the most probable sequence of phonemes and its time intervals. This method was tested with two different contexts; a free phone-loop context and a bigram context, described below.

3.2.1.1 The Phoneme Recognizer with Free Phone-loop

A Free Phone-loop can relate itself to a unigram context, that defines an item dependent at its own, in this case meaning that an occurrence of a phoneme depends only of itself and the probability a phoneme occur after a phoneme is the same as any other one. Figure 5 represents the free-phone-loop used in this method, where the links possess the transition probabilities.

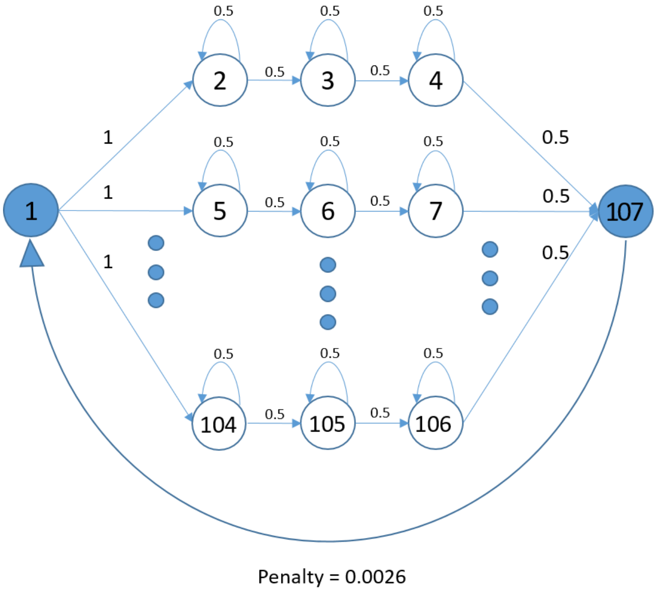

Figure 5 – Universal background model adjusted for a 35 phonemes network with free-phone-loop

3.2.1.2 The Phoneme Recognizer with Bigram context

In a bigram context a phoneme probability is conditioned by the preceding phoneme. Basically, a bigram model acknowledges the fact that after an occurrence of a phoneme some phonemes are more likely to succeed than others.

For example, in the Portuguese language the phoneme "a" is highly unlikely, if not impossible, to be followed by another "a", however is more likely to be followed by the phoneme "r" like in the word "caro", see Appendix A.

3.2.1.3 Results

Taking into consideration the work developed before (Costa, 2015) it is safe to assume that a bigram context will produce better results in terms of phoneme recognition, however to verify which recognition method is better a simple test using the Levenshtein Distance was used. For two given strings, the Levenshtein Distance calculates the minimum single-character edits (that can be insertions, deletions or a substitution) required to change a word or a phoneme into the other resulting in difference between the two sequences.

This algorithm is applied to the pair of sequence of phonemes: the reference one (ground truth) and the sequence that is given by the phoneme recognizer. Before computing the phoneme error (PER) rate, the two sequences must be best aligned in order to count the phonemes that were inserted, deleted, substituted or well recognized. Then the PER is computed as:

$$PER = \frac{S+D+I}{N_{Ref}} \tag{3.1}$$

where S represents the number of substitutions, D the number of deletions and I the number of insertions of the recognized phonemes related to the reference phonemes. The sum of substitutions, deletions and insertions is divided by the number of the phonemes in the reference transcription, N_{ref}, resulting in the Phoneme Error Rate (PER).

In this case, to confirm the assumption that bigram context produces better results than free-phone-loop it was considered 2268 utterances, present in the manual annotation.

For each network (35, 40 or 41 phonemes) a Phoneme Error Rate (PER) between the expected sequence of phonemes and the sequence of phonemes recognized was calculated for both bigram and free-phone-loop, using the tool editdistance with insertions, deletions and substitutions weighted as 1.

Table 1 provides the obtained results.

	Unigram	*Bigrama*
	PER	*PER*
35 phonemes network	*23.48%*	*22.51%*
40 phonemes network	*24.97%*	*23.98%*
41 phonemes network	*23.91%*	*23.05%*

Table 1 – PER of the phoneme recognizer for bigram and free-phone-loop

Results show that for all networks a bigram context possesses a PER around 23%, which is about 1% less than the PER of free-phone-loop for each network.

As expected a bigram context produced slightly better results than a unigram context which confirms the proposed assumption.

In light of the results obtained no further testing was considered and for feature extraction the bigram context was chosen.

Notice that even though the network architecture with 35 phonemes, described in 2.2.3, possesses better results in terms of PER than the other 2 networks, due to the fact that bellow in the Word-Alignment the better network is the one trained with 40 phonemes and the results expressed here are quite similar for the 3 networks.

3.2.2 The Word-Alignment method

The second method for Posteriorgram decoding was a Word-Alignment. This method consists in a "forced alignment" of a sentence, that is assuming that the uttered words correspond exactly to the words in the prompt sentence.

The method takes a given sequence of words and finds the most likely position for each word allowing pauses (or noise, respiration or other events - model “oth”) between words. Basically when a child reads two or more words without any pauses between them, the method will pass from one word to the next one without marking a pause event.

For better understanding, figure 6 represents a direct acyclic graph which is representation of the task-grammar model used by this method.

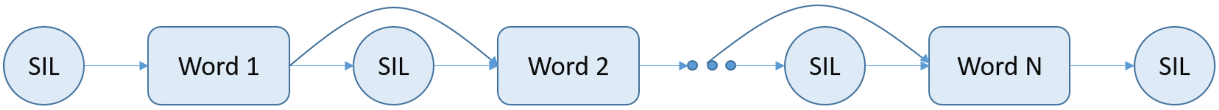

Figure 6 –Acyclic graph of the task-grammar without Garbage Phone

This method needs a phonetic transcription of each word in the sentence.

Three dictionaries with a phonetic transcription of words were used for that matter, one containing 51072 words, another containing 2753 words and a specific one with 524 pseudo words, the last two dictionaries are specifically from the LetsRead project.

Figure 7 shows a simple example of the response of this method for a sentence with two words, W1 and W2. W1 models the first word in the sentence while W2 models the second word in the sentence. From the first frame until the frame t_1 there was silence (pause), while from t_1 until t_2 the first word was pronounced and from t_2 until t_3 the second word was pronounced without a pause between them.

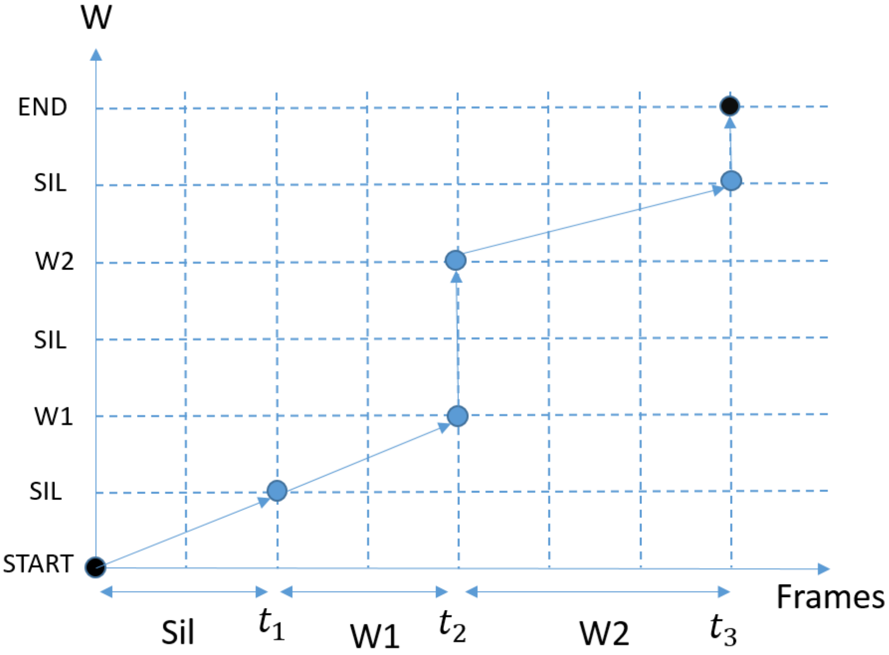

Figure 7 – Word-alignment with graph model of figure 6, redesigned from (Lopes, 2011)

Two variations of this method were taken into consideration. The first variation one only allows silences between words while the second variation allows an addition of a Garbage Phoneme, described below, intending to consume frames when a PRE or a REP evens are present in the utterance.

3.2.2.1 Word-Alignment with Garbage Phone

The second variation of the Word-Alignment used is a model that allows the model "oth" in parallel with "sil".

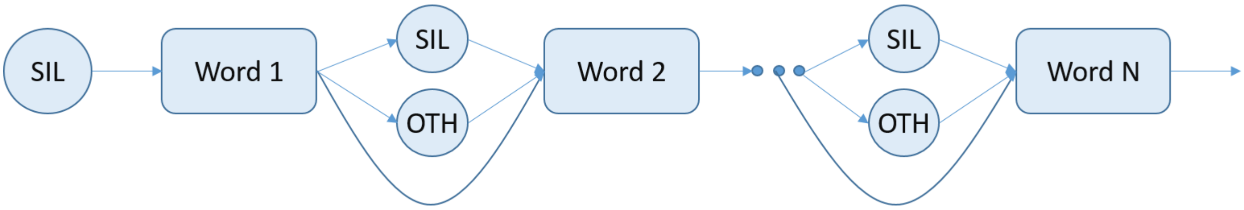

Figure 8 – Acyclic graph of the task-grammar with Garbage Phone

Figure 8 represents a direct acyclic graph which is a representation of the task-grammar model used by this method, and as we can observe both "sil" and "oth" models are in parallel and with an epsilon-link that allows the method to skip this two models if more advantageous.

The strategy is to attribute a probability value for all frames of the model "oth", which will be designated as "Garbage Phoneme".

The term Garbage Phoneme is due to the fact that between words this model would consume the frames related to some of the children's disfluencies like a pre-hesitation or a word repetition, PRE and REP respectively.

In this case the sum of the probabilities in each frame of the Posteriorgram is no longer one.

The best probability value for the model "oth", the Garbage Phoneme, was obtained by attempt and error, meaning that perhaps there is a better value than the one obtained. In order to find this value a few tests were made described below in results.

3.2.2.2 Results

The first step was to determine which strategy produces better results in terms of word alignment. This step, using both strategies with and without *Garbage Phone* and for different probability value of the *Garbage Phone* expressed in the phoneme model *"oth"*.

It was considered 450 utterances with words manually annotated corresponding to the reference annotation.

It was then calculated for each strategy the percentage of well aligned words for different percentages of interception between the word-alignment method results and the ones annotated manually. Basically, each aligned word time marks was compared with the reference ones. A word is considered well aligned according to a percentage of interception. The follow example clarifies the method.

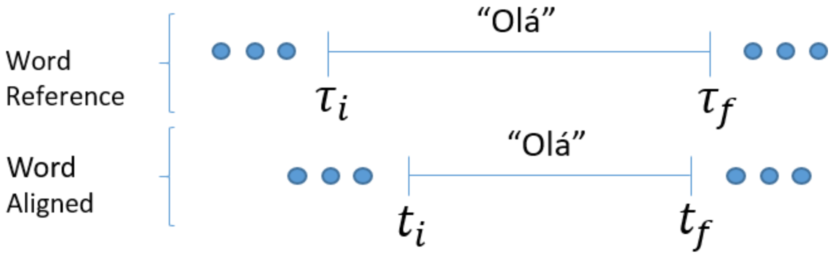

Figure 9 – Example of a word reference aligned with a word from the word-alignment

Represented in figure 9 there is an example of a word alignment with the word reference, in which τ_f and τ_i corresponds to the final and the initial time, respectively, of the reference word, while t_f and t_i corresponds to the final and the initial time, respectively, of the aligned word.

The procedure to calculate the interception percentage of two tags was the following. First the frame interception between tags is calculated (3.2):

$$T = \min(t_f, \tau_f) - \max(t_i, \tau_i) \tag{3.2}$$

Then the percentage of interception of the reference tag in the aligned tag is calculated (3.3):

$$P_1 = \frac{T}{\tau_f - \tau_i} \tag{3.3}$$

The same as in (3.3) but now to calculate the percentage of interception of the align tag in the reference one (3.4):

$$P_2 = \frac{T}{t_f - t_i} \tag{3.4}$$

At last, the percentage of interception between tags is given by the smaller one (3.5):

$$P = \min(P_1, P_2) \tag{3.5}$$

This value, *P* (3.5), was used in order to determine the best method in terms of word alignment. Figure 10 compares the world-alignment with and without the strategy of “Garbage Phoneme”, calculating the percentage of well aligned words, in comparison with the reference, for different values of *P* (*P*=50% until *P*=95%).

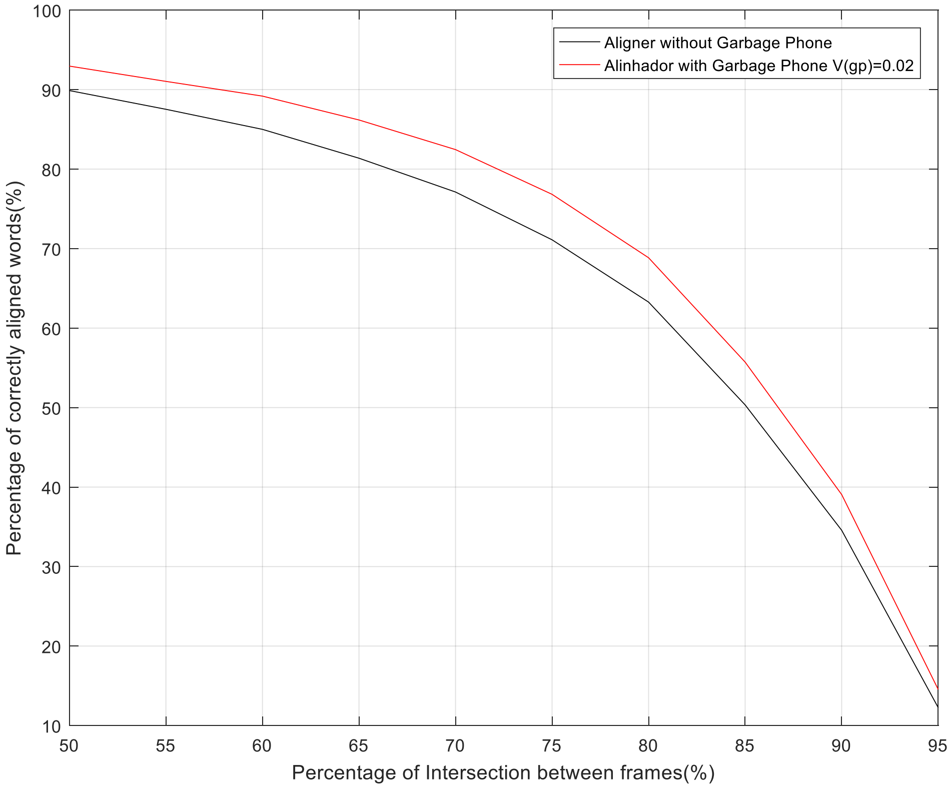

Figure 10 – Align percentage, with Garbage Phoneme versus Garbage Phoneme (35 phoneme network)

Figure 10 expresses the advantage in terms of word alignment of the word-alignment with a Garbage Phoneme over the word-alignment without Garbage Phoneme model, considered for the 35 phonemes network.

In this step several probability values for the Garbage Phoneme were considered ranging from 1^{-10} to 0.8 and was determine that the optimal value would be between 0.01 – 0.1.

The same procedure was used for the other 2 networks (40 and 41 phonemes). Since the results were quite similar, from a graphical representation are difficult to determine which network has the best results. So a mean value of the respective line, like the one presented in figure 10, was calculated where the percentage of well aligned words were divided by 10, the number of intervals.

Table 2 shows the results obtained, when attributing a probability value to the Garbage Phone of 0.02 and 1 divided by the number of phonemes in the respective network.

Assigned value for Garbage Phone	*35 Phonemes Network*	*40 Phonemes Network*	*41 Phonemes Network*
0.02	69.68%	72.82%	67.80%
$\frac{1}{\#phonemes}$	69.56%	72.80%	68.44%

Table 2 - Mean Interception of labels with the reference for all networks

As shown in the table 2, the network with 40 phonemes has the better mean percentage of well aligned words from the word-alignment with a value around 72.8%.

3.3 Detecting Well Pronounced Words

For a given utterance it is necessary to differentiate which words were well pronounced from those with disfluencies. The methods described before will be used to detect well pronounced words by resorting to a hypothesis test.

The null hypothesis H0 is correctly pronounced word. A true positive is when the system correctly accepts H0. A true negative is the when the system correctly rejects H0. A false positive or false alarm is when the system incorrectly accepts H0. A false negative or miss is when the system incorrectly rejects H0. Basically when a child pronounces correctly a word a true positive occurs if the method accepts it as well pronounced and a false negative occurs if the method rejects the word as well pronounced. When a child badly pronounces a word a true negative occurs if the method rejects it as well pronounced and a false positive occurs if the method accepts the word as well pronounced.

Using 1200 manually annotated utterances as the ground truth and the log-likelihood (score of word) given by the word-alignment as the decision measure. The hypothesis H0 is accepted when the word score is higher than a given threshold and rejected when is lower.

Other measures besides the word score were considered.

The measure S1 (3.6), which is the score of the word w divided by the number of phonemes (#phonemes(w)) that the word *w* contains,

$$S1(w) = \frac{Score(w)}{\#phonemes(w)} \qquad 3.6$$

The measure S1 (3.7), which is the score of the word w normalized by the word interval given by the word-alignment (#frames(w)).

$$S2(w) = \frac{Score(w)}{\#frames(w)} \qquad 3.7$$

The measure LLR1 (3.8) is a log-likelihood ratio between the log-likelihood (*score(w)*) of a word given by the word-alignment and the log-likelihood of the recognized phonemes given by the phoneme recognizer (score_phonemes(n)) in the respective word interval.

$$LLR1(w) = -abs(score(w) - (\sum_{n=1}^{N} score\_phonemes(n))) \qquad 3.8$$

The measure LLR2 (3.9) is a normalization of LLR1 by the word (w) interval given by the word-alignment.

$$LLR2(w) = \frac{LLR1(w)}{\#frames} \qquad 3.9$$

A detection Error Tradeoff (DET) is used to define the measure to use to test the hypothesis H0 and the respective threshold. A detection Error Tradeoff is a comparison between the false positive rate and the false negative rate at each threshold. The Equal Error Rate (EER) is the point in the DET where false positive rate and false negative rate are more similar (Martin et al,1997).

Since the 40 phonemes network produced better results is the only presented results here for this network.

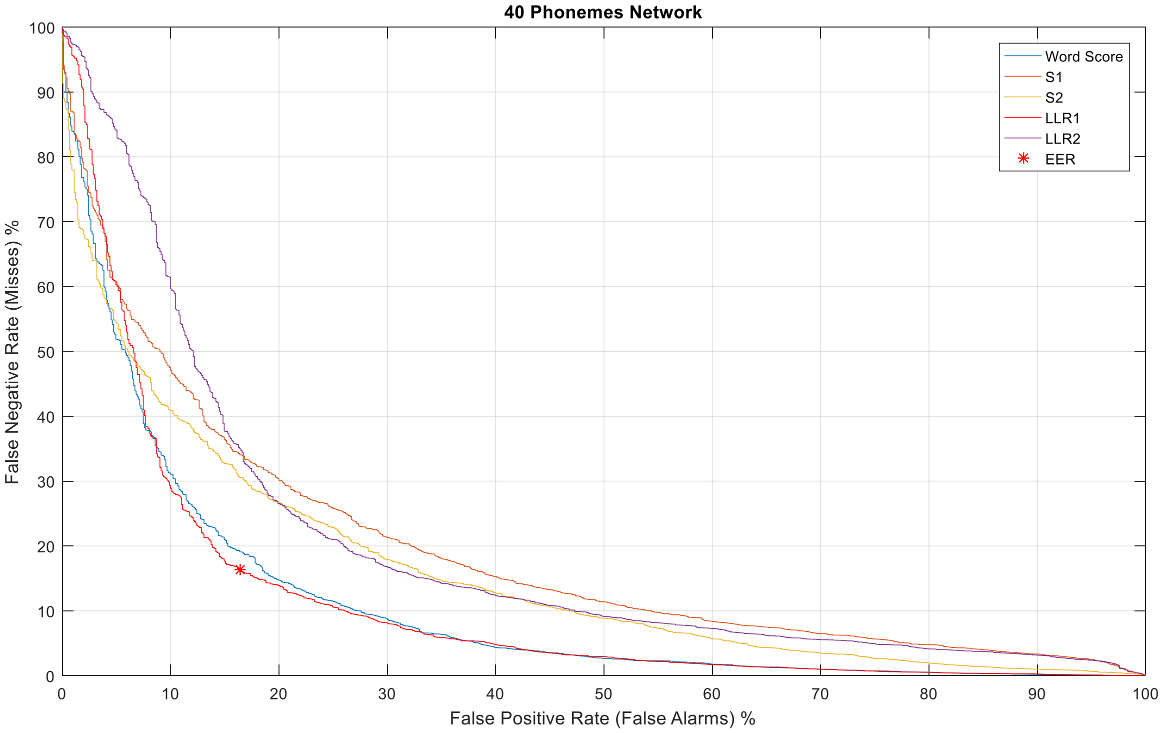

Figure 11 – Det (3.6-3.9) for 1200 utterances (40 phonemes network)

As expected LLR1, as shown in figure 11, produced the best results with an Equal Error Rate, of 16.38% False Negatives (Misses) and 16.37% of False Positives (False Alarms), setting the threshold at -26.90 since the Equal Error Rate seems the better measure to decide the hypothesis H0. The word score by itself produced an unexpected result, in the positive way, however this metric it is not so trustworthy for smaller words. Notice that for the lowe False Positive Rate the metric S2 should be considered.

4 Assessing a Child’s ability to read

This chapter focusses in describing the chosen features for evaluating a children’s reading ability and the proposed models for this purpose.

All the features were extracted from both methods described before in chapter 3.

4.1 Features for Reading Performance

The proposed features pretend to withdraw some characteristics from the children’s related to their ability to read utterances.

A commonly index to assess the oral reading fluency of a child is *Words Correct Per Minute (WCPM) (bolaños et al, 2011).* Even though this feature is well documented as a good indicator towards accuracy and rate of a child’s reading ability, it does not measure the third component of Oral Reading Fluency, which is the expressiveness of the child. There is also another problem concerning WCPM, which is the uncertainty in the decision of defining words well pronounced (correct words), using the method defined in the previous chapter3. Another set of features should be considered in order to produce a more robust analysis of the child’s speech.

A set of 35 features were derived from the two methods referred to before in 3.2. The goal is that these features will correlate well with the level of fluency and reading ability of the children, as indicated by the teachers. This set of features can be inserted into two different categories. One category relates itself with the lexical properties of the utterance, while the other category relates itself with the prosodic properties of the utterance.

Lexical features are those related to the words, or extracted from a word-based level of the utterance itself.

Prosodic features are those related with speech behaviors, those elements of a speech that are not individual phonetic segments but instead properties of the speech itself.

For example, Words Correct Per Minute is a lexical feature, however the mean duration of correct words is a prosodic feature.

From the set of 35 features, 20 are considered as lexical features while 15 are considered as prosodic features.

Some features were taken exclusively from the results of the automatic speech recognition methods, Phoneme Recognizer and Word-Alignment. Some were withdrawn exclusive from the Word-Aligner method while others were obtained due to a mixture of both methods. After some experiments, 10 features were taken from the aligner method; 16 from the Phoneme-Recognizer method and 9 from the mixture of both methods.

Some features behave basically the same way, providing most of the times the same results, however both features from the Word-Aligner and the Phoneme-Recognizer method need to be considered.

Lexical Features:

Features taken with the Phoneme-Recognizer, based on all the recordings for evaluation of one child:

F1. Number of recognized phonemes per minute (the number of phonemes output by the recognizer divided by the duration of the utterance, times 60). It is expected that a child with higher number of recognized phonemes per minute possesses a better ability to read.

F2. Mean of Log-Likelihood (phoneme scores) of the recognized phonemes normalized per the total number of recognized phonemes.

F3. Mean of Log-Likelihood (phoneme scores) of the recognized phonemes normalized by the number of frames (of 10 milliseconds).

F4. Mean number of recognized phonemes per word.

F5. Mean Levenshtein distance (edit distance) per word. Number of insertions, deletions and substitutions between the recognize phonemes and the phonemes in the reference words.

F6. Mean Levenshtein distance (edit distance altered as described below) per word.

F7. Mean Levenshtein distance (edit distance) per frame. Number of insertions, deletions and substitutions between the recognize phonemes and the phonemes in the reference words.

F8. Mean Levenshtein distance (edit distance altered as described below) per frame.

F9. Difference between the number of recognized phonemes and the total number of phonemes in the sentence.

F10. Difference between the number of recognized phonemes and the total number of phonemes in the sentence normalized per frame (10 milliseconds) of the spoken regions (regions belonging to phonemes, non-silences).

F11. Number of phonemes in the sentences divided by the total number of recognized phonemes.

Features taken with the Word-Aligner.

All the measures are, based on the whole set of recordings for evaluation of the target child. The features are the following:

F12. Number of aligned words per minute.

F13. Mean of Log-Likelihood (word scores) of the words at the output of the aligner divided by the number of reference words.

F14. Mean of Log-Likelihood (word scores) of the words at the output of the aligner divided by the total number of frames of the aligned words..

Mixture between the two methods:

F15. Total Log-Likelihood Ratio (LLR1) between the score obtained by the word in the Aligner method and the score obtained by the recognizer method, normalized per frame (10 milliseconds) of the word regions in given by the Aligner method.

F16. Total Log-Likelihood ratio (LLR1) between the score obtained by the word in the Aligner method and the score obtained by the recognizer method, normalized per frame (10 milliseconds) of the spoken regions given by the Recognizer method.

F17. Total Log-Likelihood ratio (LLR2) between the score obtained by the word in the Aligner method and the score obtained by the recognizer method, normalized per phoneme relative to the sentence.

F18. The quadratic mean of the total Log-Likelihood ratio between the score obtained by the word in the Aligner method and the score obtained by the recognizer method, normalized per word (3.8).

$$F18 = -rms\left(\sum_{n=1}^{\#words} LLR1\,(n)\right) \qquad 4.1$$

F19. The quadratic mean of the total Log-Likelihood ratio between the score obtained by the word in the Aligner method and the score obtained by the recognizer method, normalized per frame (10 milliseconds) of the spoken regions given by the Recognizer method (3.8).

$$F19 = -rms\left(\sum_{n=1}^{\#words} LLR2\,(n)\right) \qquad 4.2$$

F20. Words Correct Per Minute. This feature is determined using the metric LLR1 when the decision of accepting correct words is taken using a threshold on LLR1 of -26.9.

Features F6 and F8 uses a changed editdistance (Levensthein Distance) algorithm, in which insertions and deletions have a weight of 8 (instead 1 as before), and for a substitution the weight depends on the phoneme that is substituted, according to a phonetic proximity that changes from 3 to 9. The table of distance between phonemes is embedded in the implementation of the edit-distance algorithm.

Prosodic Features:

The prosodic features are based on the fact that an utterance has pauses (silences) in the beginning and optionally between words or even intra-words that can be separated from speech intervals. The final silence is discarded in all measures involving pauses because it has no meaning for reading performance.

Phoneme-Recognizer Method:

F21. Number of pauses (silences bigger than 50 milliseconds) per minute (the total duration of the SIL phonemes on the phoneme recognizer output).

F22. Mean duration of phones (mean duration of the total recognized phonemes).

F23. Man of the durations of speech intervals (discarding silences) divided by the number of reference words.

F24. Percentage of the silence duration relative to the total utterance duration.

F25. Mean duration of the initial silence (reaction time to the prompt).

F26. Percentage of the initial silence duration relative to the total utterance duration.

F27. Duration of the biggest silence region discarding the first silence (maximum hesitation).

F28. Maximum duration of a recognized phoneme (the longest phone in the child's utterances).

Word-Aligner Method:

F29. Number of pauses per minute (the number of silence or respiration events longer than 50 milliseconds divided by the duration of the utterances - times 60).

F30. Mean duration time of words normalized by the number of phonemes in it.

F31. Mean duration time of a word.

F32. Percentage of the silence duration relative to the total utterance duration (same as F24 but taken from the word-aligner output).

F33. Mean duration of the initial silence

F34. Percentage of the initial silence duration relative to the total utterance duration (same as F26 but taken from the word-aligner output).

F35. Average number of words that have no pauses between them.

4.2 Model for assessing children's reading ability

This section discusses how to generate a robust model to assess the children's reading ability, resorting to the proposed features in 4.1. Several methods of regression analysis were considered in order to estimate a relation between the proposed features and the children's indexes attributed by the scorers (teachers), such as Support Vector Machine, Multi-Layer Perceptron, Linear Regression and others. Using an open-source software for data-mining (WEKA 3) quickly was realized that a Multivariate Linear Regression not only produces as good results as other methods, as it is a much simpler model and easier to implement (in this case using Matlab).

Several different approaches were tried, not only in order to evaluate a child's reading ability but also to understand how important the different characteristics of a child's speech are.

This analysis allowed to determine which proposed features are more relevant in a child's speech than others.

The features that produced the results presented below were obtained using 1200 utterances from 150 children as described in 3.1.2. Only 1200 utterances were used, even though the database possesses more utterances of the 150 children, because only this 1200 utterances were evaluated by the teachers.

In the following approaches, the linear models used will attempt to adjust the numerous elected features to a ground-truth, which is the index provided by the teachers to each of the considered 150 children. The index used as ground-truth is the one explained in 3.1.2 and labelled as *z-norm index*, which is basically a normalization per evaluator of the raw medium indexes given by the teachers.

4.2.1.1 First Approach – Evaluation of each feature

In order to evaluate each of the proposed features in terms of assessing a child's reading ability, a linear regression model was applied to each of the features individually using a total of 1200 utterances from 150 different children ranging from first to fourth grade. 80 % of the utterances was used for training while 20% of the utterance was used for testing - resulting in 120 children for training and 30 children for testing.

The results in the table 3 for the proposed features are merely an indicator of which of those features would be excellent and which of those would be bad in order to assess a child's reading ability.

Features	*ρ*	*Features*	*ρ*	*Features*	*ρ*	*Features*	*ρ*
F1	0.9298	*F11*	0.8924	*F21*	0.7960	*F31*	0.7959
F2	0.8305	*F12*	0.9452	*F22*	0.8662	*F32*	0.8699
F3	0.2526	*F13*	0.8104	*F23*	0.7656	*F33*	0.7554
F4	0.7981	*F14*	0.5805	*F24*	0.8416	*F34*	-0.0901
F5	0.5181	*F15*	0.6247	*F25*	0.6935	*F35*	0.8134
F6	0.5045	*F16*	0.5699	*F26*	-0.1628		
F7	0.6369	*F17*	0.7224	*F27*	0.7044		
F8	0.6405	*F18*	0.7723	*F28*	0.6871		
F9	0.8668	*F19*	0.3899	*F29*	0.5079		
F10	0.8525	*F20*	0.9336	*F30*	0.8619		

Table 3 – Correlation Coeficient (ρ) for each individual feature with z-norm index

By analyzing the results in table 3 some features correlated well with the child's reading index as expected, such as the features F1 (Number of recognized phonemes per minute) F12 (Number of spoken words per minute) F20 (Words Correct Per Minute), marked in green. These features were expected to provide good results, however once again this three features are intrinsically related only with the rate and accuracy in a child's oral reading fluency. None of these are features of a prosodic nature, however there is some interesting prosodic features that provided an excellent correlation with the children's reading index, obtaining a correlation coefficient (*ρ*) around 0.85, such as the features F22 (Medium duration of a recognized phoneme), F24 (Percentage of the silence areas in relation to the total utterance time), F29 (Number of pauses (silence or respiration region bigger than 50 milliseconds) per minute), F32 (Medium duration of the initial silence area) and F10().

While some of the proposed features produced good results in terms of correlation with the child's reading index others features produced awful results in terms of correlation between them and the child's reading index. The features marked in red in the table 3 identify these preposterous features which came as a supervise due the fact that was expected a much better correlation between these features and the children's reading index.

The features F26 (Percentage of the initial silence area duration relative to the total utterance time) and F34 (Percentage of the initial silence area duration relative to the total utterance time), obtained with the Recognizer method and the Aligner method respectively, resulted in a negative PPMCC near zero which means there is almost no correlation what so ever with these features and the child's reading index. The features F3 (Medium of Log-Likelihood Ratio (phoneme score) of the recognized phonemes, normalized per the total number of recognized phonemes), also marked in red in table 3, and the feature F19 (The quadratic mean of the total Log-Likelihood ratio between the score obtained by the word in the Aligner method and the score obtained by the recognizer method, normalized per frame (10 milliseconds) of the spoken regions given by the Recognizer method) resulted in an awful PPMCC even under 0.5.

These awful results of the "bad" features (F26, F34, F3, F19) referred to before, comes as a surprise due to the fact that was expected much better results, hereupon this four features are ruled out in any further examination.

4.2.1.2 Second Approach – Evaluation of the Lexical Features versus the Prosodic Features

As described before, the lexical characteristics of a children's speech provide a good analysis in terms of rate and accuracy in reading fluently, however expressiveness is also a good indicator of a fluent reader and this is more related to prosody of speech.

This section intends to determine the response of both Lexical and Prosodic Features to the reading ability of a child.

Once again, and as before in the first approach, the data set of features related to 150 children's is divided in 80% (120 children) for training and 20% (30 children) for testing. Keep in mind that 4 features (F26, F34, F3, F19) were discarded resulting in a total of 31 features, where 18 are lexical related features and 13 are prosodic related features.

Table 4 shows the results using a linear regression with the dataset as described adjusting to the *z-norm indexes*.

Lexical Features		*Prosodic Features*	
ρ *(training set)*	*0.9507*	ρ *(training set)*	*0.9365*
ρ *(testing set)*	*0.9413*	ρ *(testing set)*	*0.9302*
RMSE	*0.3657*	*RMSE*	*0.4136*

Table 4 – Comparison of the linear model for only lexical or prosodic features

As expect, lexical-based features produce slightly better results with a correlation coefficient of 0.9413 and an associated Root Mean Square Error (RMSE) of 0.3657, however prosodic features can also be a good metric to assess a child's reading ability, since it resulted in 0.9302 of correlation coefficient and an associated error of 0.4136.

4.2.1.3 Third Approach – The Proposed Models

In this section, two models for assessing a child's reading ability are proposed using some of the features described before. In this section it is used the technique of "k-fold cross validation" in order to certify and strengthen the proposed model, as explained below.

In all the experiments a k=5-fold cross validation was used, portioning the dataset of 150 children into partitions of 120 children for training and 30 children for testing, rotating the data.

For better understating of this technique, figure 12 represents an example of a 5-fold cross validation.

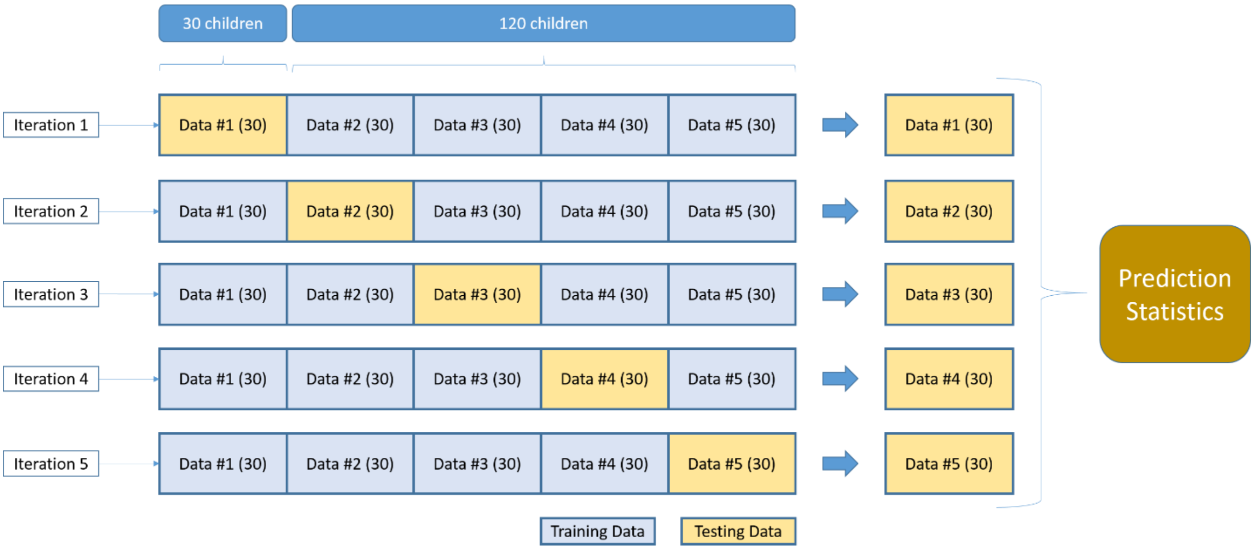

Figure 12 – k-fold cross validation in application for this work. The yellow blocks represent data for testing and blue blocks for training. The final result is the average of the 5 partial results.

Proposed Model 1 – All Features

First step was idealizing a multivariate linear regression model to assess children's reading ability containing all features (excluding once again features F26, F34, F3 and F19), in order to be used as a baseline and as comparison towards the final proposed model. This model by itself generates interesting results, presented in Table 5, since for the testing dataset of the 5-fold cross validation technique generated a mean of the correlation coefficient values of 0.9396 between the children's reading index predicted by the model and the z-norm index.

Proposed model 2 – Simplified Model

The second proposed model is a S*implified Model* containing only 3 features. The idea behind this model is to assign for each of the three bases of Oral Reading Fluency - rate, accuracy and expressiveness - a feature that by itself would proper characterize each of this three components. The three features were chosen using a Matlab tool called `stepwiselm`, which produces a linear regression model by constantly using forward and backward stepwise regression that picks the best features providing a regression model with only a few features. In each iteration of the k-fold cross validation this Matlab tool chose different features, which were the followed:

First iteration: F1, F11, F16, F35, F1*F16

Second iteration: F1, F11, F5, F16, F35

Third iteration: F1, F11, F16, F17, F35

Fourth iteration: F1, F11, F20, F27

Fifth iteration: F1, F11, F20, F31, F35, F1*F35

With this features as reference several tests were made in order to find the three features that best express each individual component of the Oral Reading Fluency.

In terms of children's reading accuracy, feature F11 (Words Correct Per Minute) was chosen. In terms of child's reading rate feature F1 (Recognized Phonemes Per Minute) was chosen. Even though the features complement one another, these are both good indicators of a child's rate and accuracy. In terms of child's reading expressiveness the feature F35 (Average number of words between silence regions) was chosen, since it is anticipated that a more expressive child read more words without pausing.

Once again, as before, the data set of features related to 150 children's was subjected to a 5-fold cross validation using exclusively the data from the features F1, F11 and F35 in order to evaluate to the quality of the linear regression model.

The results for this S*implified Mode*, expressed in table 5, provides an evident good correlation between this model with the children's reading ability, since the mean of the correlation coefficient is 0.9440. See Appendix B to use this model.

Proposed model 3 – Final Model

The third (Final Model) is the more robust proposed model to assess a child's reading ability and was obtained after several experiments. Using the features referred before generated by the Matlab tool `stepwiselm`, crossing almost all the features with each other and also by "trial and error", The better result outcome 2 more features, which are a combination of three of the already existing features, F1, F16 and F35.

The first newly created feature, which will be designated as FE1, is a multiplication between F1 and F16 and the second newly created feature, which will be designated as FE2, is a multiplication between F1 and F35. The model with this features additions continues to be a linear regression model.

This model was generated using the followed features F1, F11, F17, F30, F32, F35, FE1, FE2 and the results of the model are expressed in the table 5. See Appendix B to use this model.

5-fold cross validation (mean values)	*ρ (training set)*	*ρ (testing set)*	*RMSE*
All Features	*0.9587*	*0.9396*	*0.3856*
Simplified Model (3 features)	*0.9412*	*0.9440*	*0.3920*
Final Model	*0.9524*	*0.9520*	*0.3535*

Table 5 – 5-fold cross validation mean values for proposed models (all features, simplified model and final model)

The results for the Final *Model*, expressed in table 5, provides an evident excellent correlation between the predicted reading indexes with the children's reading ability, since the mean of the correlation coefficient is 0.9520. The Final Model produced, as expected, better results comparing with the All Features Model and the Simplified Model, with an improvement on both the correlation coefficient and the root mean square error.

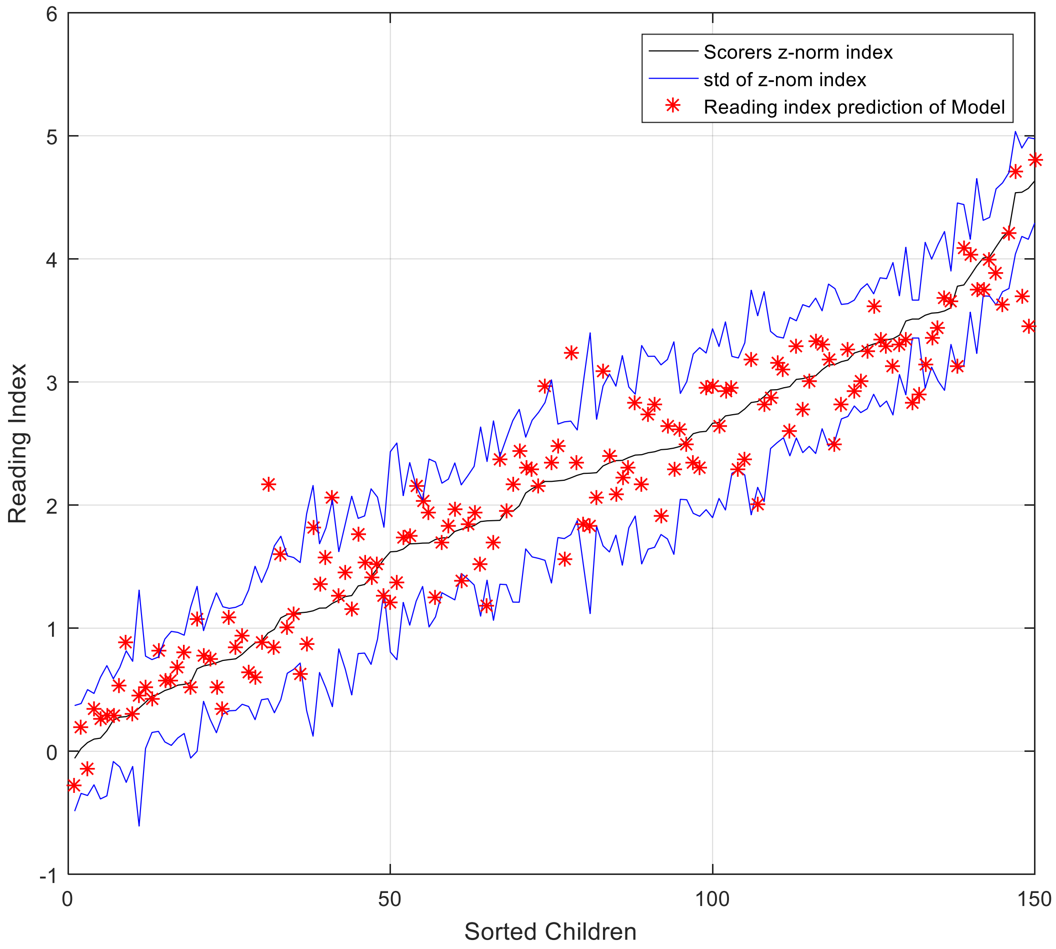

Figure 13 – Predicted indexes by the Proposed Model vs z-norm teachers indexes ± standard deviation

Figure 13 provides us with a visual representation of the Proposed Model 3 (The final Model) behavior in providing a child with a reading ability index. As is visible in figure 13, almost all the predicted reading indexes of each child are nearby the z-norm index provided by the teachers or at least inside the standard deviation associated with the z-norm index provided by the teachers. From the 150 predicted reading ability values by the Proposed Model only 20 are outside of the z-norm standard deviation region, and even these 20 values are close to this region by analyzing the figure 13.

For further validating of the final model as a robust model to assess a child's reading ability a comparison between the teacher's standard deviation and the predicted index error is expressed in figure 14.

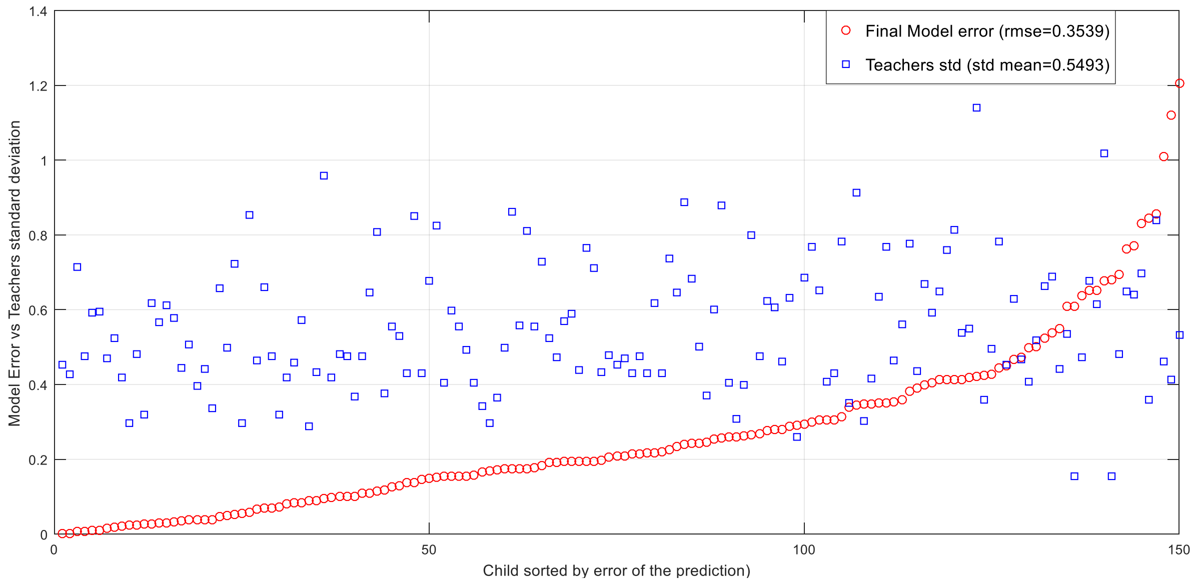

Figure 14 – Teachers Standard deviation vs Proposed Model Error

The final model produced in average a smaller difference between the error of the predicted indexes and the z-norm indexes than the standard deviation of the teachers himself.

The higher teacher's deviation towards the z-norm index is due to the subjective nature of the task (Narayanan).

The Proposed Model to assess children's reading ability was designed based on a reading task of 5 sentences and 5 pseudo-words. and in order for a good assessment value it is desirable that a similar reading task is employed.

4.3 System Examples

In this section, 4 examples, with 4 different children, using the system described until this moment are presented. The fourth grade child example uses an utterance of pseudowords.

First example, presented in figure 15, a child from the first grade was chosen. The read utterance is the following: “era uma vez um elefante muito pequenino e muito enfezado.”

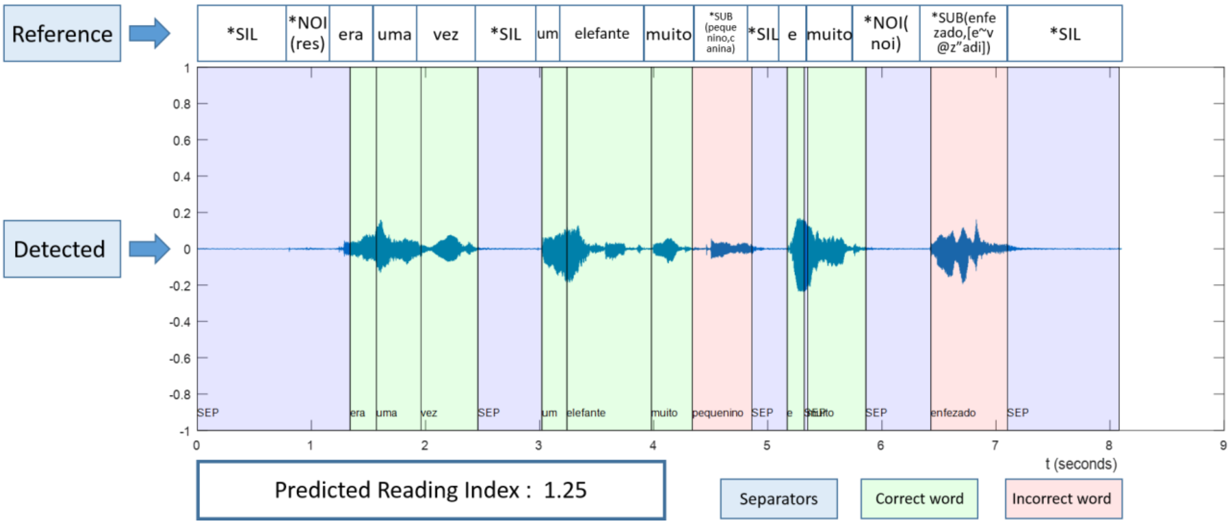

Figure 15 – Example of the model for a first grade with an overall predicted index

Two of the 10 words read by the child were badly pronounced, “pequenino” and “enfezado”. The system correctly predicted the well pronounced words as the badly pronounced ones, as correctly aligned the words, with small deviations in the time marks.

The predicted reading index presented is the overall reading index of this child, which is 1.25. The reading index attributed by the teachers is 1.721 ± 0.629. In this case the proposed model assesses this child closely to the index of the teachers and inside the standard deviation of the index.

The second example, presented in figure 16, corresponds to a child from the second grade and the respective utterance is the following: “Há cinema e circo na cidade”.

This utterance possesses 6 words which were all correctly pronounced.

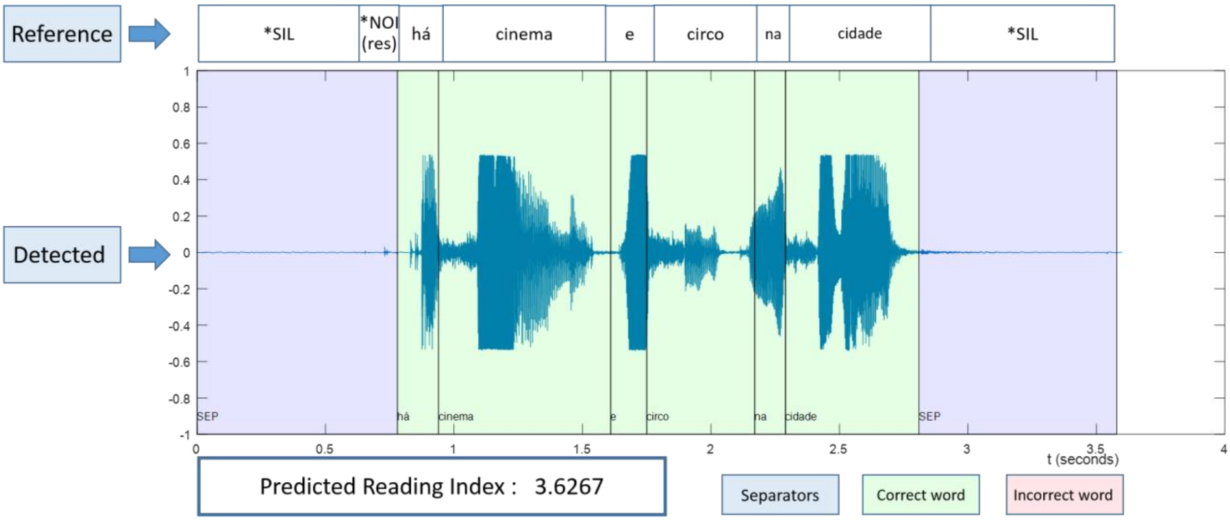

Figure 16 – Example of the model for a second grade child with an overall predicted index

The system correctly accepted 6 out of the 6 words as well pronounced and aligned 6 out of the 6 words correctly.

The predicted reading index presented is the overall reading index of this child, which is 3.6267. The reading index attributed by the teachers is 4.175 ± 0.442. In this case the proposed model assesses this child closely to the index of the teachers but outside of the standard deviation of the index.

The third example, presented in figure 17, corresponds to a child from the third grade and the respective utterance is the following: “O Manuel se não os visse não os podia apalpar”.

This utterance possesses 10 words which were all correctly pronounced however the child repeated the words “se não visse”.

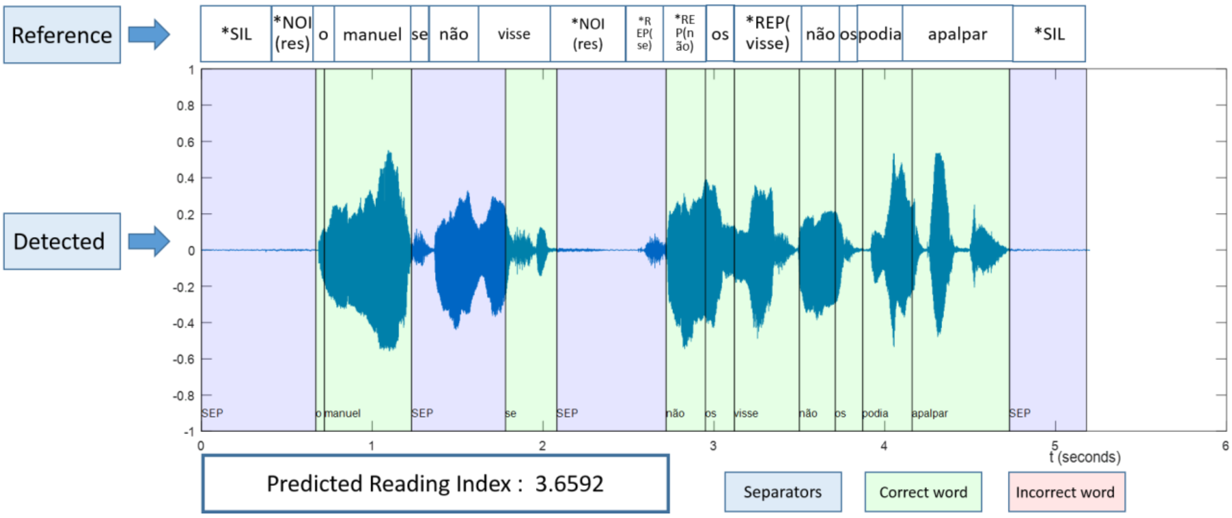

Figure 17 – Example of the model for a third grade child with an overall predicted index

The system correctly accepted 10 out of the 10 words as well pronounced and aligned 9 out of the 10 words correctly.

The predicted reading index presented is the overall reading index of this child, which is 3.6592. The reading index attributed by the teachers is 3.603 ± 0.298. In this case the proposed model assesses this child closely to the index of the teachers and inside the standard deviation of the index.

The fourth example, presented in figure 18, is a child from the fourth grade, however in this example the respective utterance is a sentence with 5 pseudo-words which is the following: “zecla simbrelhos luga grefão hoberem”.

This utterance possesses 5 pseudo-words in which 4 pseudo-words were correctly pronounced and the last pseudo-word was miss-pronounced.

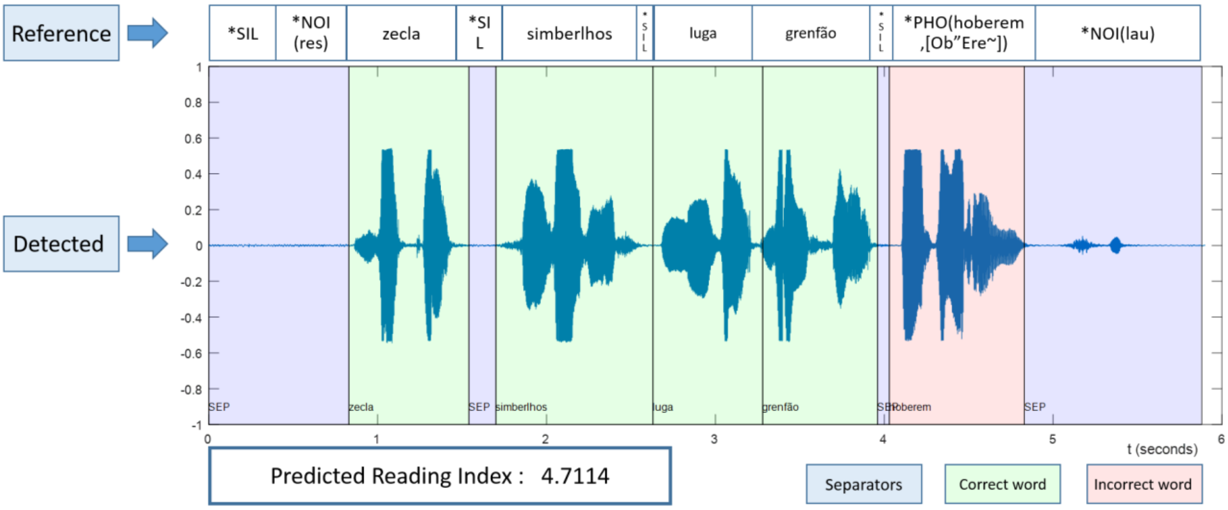

Figure 18 – Example of the model for a fourth grade child with pseudo-words

The system correctly predicted 4 out of the 5 words as well pronounced also correctly predicting the last one as badly pronounced. All the 5 words were correctly aligned, once again with a slightly deviation.

The predicted reading index presented is the overall reading index of this child, which is 4.7114. The reading index attributed by the teachers is 4.537 ± 0.4908. In this case the proposed model assesses this child closely to the index of the teachers and inside the standard deviation of the index.

5 Conclusion

The main objective of this dissertation was achieved, which was to create a system to evaluate a child's reading ability, while detecting mispronunciations and hesitations by aligning the text words with the speech of the child reading aloud.

To determine a child's reading ability index, the system recurs to a model that provides excellent results with a correlation coefficient of about 0.95 with the expert's evaluations.

In technology usually everything can be improved and so as this system. One improvement would be the use of a better performance phoneme recognizer. Other approaches can also be considered, like improving the proposed Final Model with new features, using the system for other databases, or designing it with other regression methods, etc.

This dissertation proposes some interesting results and the proposed system can be used to evaluate a child's reading ability in real environments, such as a tool aid teacher in schools.

As a personal note, it can be said that the area of speech recognition has always been an area of particular interest by the author and also a "big black box" that just became much greyer.

References

A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przyboc, “The DET Curve In Assessment Of Detection Task Performance,” Proc. Eurospeech '97, Rhodes, Greece, September 1997, Vol. 4, pp. 1899–1903.

Bolaños, D., Cole, R.A., Ward, W., Borts, E., Svirsky, E., 2011. FLORA: Fluent Oral Reading Assessment of Children’s Speech. ACM Trans. Speech Lang. Process. 7, 16:1–16:19.

Castela, Luis Miguel Bagagem, “Audio Search. Master Dissertation,” Faculty of Science and Technology – University of Coimbra, Coimbra, 2015.

Costa, Orlando Oliveira, “Development of techniques for evaluation of children's reading ability, Master Dissertation”Faculty of Science and Technology – University of Coimbra, Coimbra, 2015.

En.Wikipedia.org: Speech [Online] https://en.wikipedia.org/wiki/Speech, 2016

En.Wikipedia.org: Phonetics [Online] https://en.wikipedia.org/wiki/Phonetics, 2016

Fuchs, L.S., Fuchs, D., Hosp, M.K., Jenkins, J.R., 2001. Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis. Scientific Studies of Reading 5, 239–56.

H. C. Buescu, J. Morais, M. R. Rocha e V. F. Magalhães, “Programa e Metas Curriculares de Português do Ensino Básico,” *Ministério da Educação e Ciência,* 2015

J. Proença, O. Costa, D. Celorico, S. Candeias e F. Perdigão, “Automatic Detection of Disfluencies in Children Reading Aloud Using Task Specific Lattices,” 10th Conference on Telecommunications - CONFTELE, 2015.

J. Proença, D. Celorico, S. Candeias, C. Lopes e F. Perdigão, “Children’s Reading Aloud Performance: a Database and Automatic Detection of Disfluencies,” Conf. of the International Speech Communication Association - INTERSPEECH, 2015.

“Lets Read Project,” [Online]. Available: https://www.it.pt/Projects/Index/1938/. [acess 2016].

Logan, G.D., 1988. Toward an instance theory of automatization. Psychological Review 95, 492-

527.

Lopes, Carla Alexandra Lopes, “Enhanced Phonetic Classes in Phones Automatic Recognition, Phd Thesis” Faculty of Science and Technology – University of Coimbra, Coimbra, 2011.

National Reading Panel, 2000. 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.

Ö. G. Saracoglu e H. Altural, “Color Regeneration from Reflective Color Sensor Using an Artificial Intelligent Technique,” *Sensores,* vol. 10, nº 9, pp. 8363-8374, 2010.

Paul D. B. “Speech Rocognition Using Hidden Markov Models,” The Licoln Laboratory Journal Volume 3. Number 1 (1990).

Appendix A - Phonemes Table

Type	SPL-IT	Sampa_uc	Sampa	Examples	Phonetic Transcription
Plosives	p	p	p	presente	***p*** r @ z ë t @
	b	b	b	bola	***b*** O l 6
	t	t	t	talho	***t*** a L u
	d	d	d	dinossauro	***d*** i n O s a u r u
	k	k	k	correr	***k*** u R e r
	g	g	g	gato	***g*** a t u
Fricatives	f	f	f	fazer	***f*** 6 z e r
	v	v	v	vinho	***v*** i J u
	s	s	s	dinossauro	d i n O ***s*** a u r u
	z	z	z	fazer	***f*** 6 z e r
	S	S	S	xaile	***S*** a I l @
	Z	Z	Z	jato	***Z*** a t u
Nasals	m	m	m	mulher	m u ***L*** E r
	n	n	n	nada	***n*** a d 6
	J	J	J	vinho	v i ***J*** u
Liquid	l	l	l	bola	b O ***l*** 6
	L	L	L	mulher	m u ***L*** E r
	r	r	r	raro	R a ***r*** u
	R	R	R	raro	***R*** a r u
Vowels	i	i	i	ideia	***i*** d E ***i*** 6
	e	e	e	fazer	f 6 z ***e*** r
	E	E	E	bela	b ***E*** l 6

	a	a	a	astro	***a*** S t r u
	&		6	aquela	***6*** k E l ***6***
	O	O	O	bola	b ***O*** l 6
	o	o	o	bolu	b ***o*** l u
	u	u	u	jato	Z a t ***u***
	@	@	@	presente	p r **@** z ë t **@**
	iN	ï	i~	Inteiro	***ï*** t 6 I r u
	eN	ë	e~	Presente	p r @ z ***ë*** t @
	&N	ã	6~	coração	k u r 6 s ***ã*** ü
	oN	õ	o~	contas	k ***õ*** t 6 S
	uN	ü	u~	coração	k u r 6 s ã ***ü***

40 and 41 phonemes extra phonemes

SPL-IT	Sampa_uc	Examples	Phonetic Transcription
j	j	noite	n o j t @
jN	ü	coração	k u r 6 s ã ***ü***
w	u	jato	Z a t ***u***
wN	õ	contas	k ***õ*** t 6 S

Appendix B – Proposed Models usage

Simplified Model:

$$SMRi_{k=1} \cong -2.6519 + 0.0081 * F1 + 2.0715 * F11 + 0.1448 * F35$$

$$SMRi_{k=2} \cong -2.6972 + 0.0078 * F1 + 2.2483 * F11 + 0.1281 * F35$$

$$SMRi_{k=3} \cong -2.5492 + 0.0079 * F1 + 2.0499 * F11 + 0.1392 * F35$$

$$SMRi_{k=4} \cong -2.6111 + 0.0078 * F1 + 2.2549 * F11 + 0.1044 * F35$$

$$SMRi_{k=5} \cong -2.7910 + 0.0075 * F1 + 2.3775 * F11 + 0.1578 * F35$$

$$Ysm = mean(Ri_{k=1}, Ri_{k=2}, Ri_{k=3}, Ri_{k=4}, Ri_{k=5})$$

$$SMRI = \begin{cases} 0, & Ysm < 0 \\ 5, & Ysm > 5 \\ Ysm, & 0 \leq Ysm \leq 5 \end{cases}$$

SMRI – Simplified model predicted reading index

Final Model:

$$\begin{aligned} Ri_{k=1} \cong\ & -10.504 + 0.0203 * F1 + 3.9993 * F11 - 0.0130 * F17 + 0.0922 * F30 \\ & + 0.0399 * F32 + 0.6013 * F35 + 0.0032 * FE1 - 0.0015 * FE2 \end{aligned}$$

$$\begin{aligned} Ri_{k=2} \cong\ & -9.4076 + 0.0181 * F1 + 3.8915 * F11 - 0.0163 * F17 + 0.0784 * F30 \\ & + 0.0326 * F32 + 0.5146 * F35 + 0.0027 * FE1 - 0.0012 * FE2 \end{aligned}$$

$$\begin{aligned} Ri_{k=3} \cong\ & -9.3174 + 0.0189 * F1 + 3.8019 * F11 - 0.0217 * F17 + 0.0792 * F30 \\ & + 0.0314 * F32 + 0.5242 * F35 + 0.0041 * FE1 - 0.0013 * FE2 \end{aligned}$$

$$\begin{aligned} Ri_{k=4} \cong\ & -10.5828 + 0.0197 * F1 + 4.2882 * F11 - 0.0163 * F17 + 0.0929 * F3 \\ & + 0.0410 * F32 + 0.5303 * F35 + 0.0036 * FE1 - 0.0013 * FE2 \end{aligned}$$

$$\begin{aligned} Ri_{k=5} \cong\ & -11.2870 + 0.0206 * F1 + 4.3695 * F11 - 0.0197 * F17 + 0.1057 * F30 \\ & + 0.0404 * F32 + 0.6654 * F35 + 0.0029 * FE1 - 0.0016 * FE2 \end{aligned}$$

$$Y = mean(Ri_{k=1}, Ri_{k=2}, Ri_{k=3}, Ri_{k=4}, Ri_{k=5})$$

$$RI = \begin{cases} 0, & Y < 0 \\ 5, & Y > 5 \\ Y, & 0 \leq Y \leq 5 \end{cases}$$

RI – Final model predicted reading index