UNIVERSIDADE DE COIMBRA

Jorge Daniel Leonardo Proença

AUTOMATIC ASSESSMENT OF READING ABILITY OF CHILDREN

Tese no âmbito do doutoramento em Engenharia Electrotécnica e de Computadores, especialização em Computadores e Electrónica, orientada pelo Doutor Fernando Manuel dos Santos Perdigão e pela Doutora Sara Maria Fernandes Rato e Costa Marques Candeias e apresentada ao Departamento de Engenharia Electrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Novembro de 2018

• U C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

PhD Thesis

Automatic Assessment of Reading Ability of Children

Jorge Daniel Leonardo Proença

Advisors: Dr. Fernando Manuel dos Santos Perdigão and
Dr. Sara Maria Fernandes Rato e Costa Marques Candeias

Department of Electrical and Computer Engineering
Faculty of Sciences and Technology
University of Coimbra

March 2018

Acknowledgements

I would like to show my deepest appreciation for my advisor Fernando Perdigão, who has been my mentor for seven years. Thank you for your patience, opportunities given and guidance that made it possible for me to accomplish all of the goals I set out to achieve. I would also like to acknowledge and thank my co-advisor Sara Candeias, who was the driving force of the topic of this thesis, opened the door to international collaborations and has been a great role model of professional ambition.

I must also thank my past and present lab colleagues and project partners, Carla Lopes, Arlindo Veiga, Dirce Celorico, Pedro Olveira and many master's students, who have contributed to the success of the developed work and to a great working environment. I would also like to thank my colleagues and friends from the speech recognition community around the world, who make me appreciate being involved in this scientific area.

Andreas Stolcke and Michael Tjalve are two co-authors of recent years and have my utmost thanks for all the great ideas and guidance provided, without whom this work would surely have been poorer.

I thank my parents, family and friends for all the support given throughout this stage of my life, which is of unmeasurable importance to my happiness and well-being, and would need another thesis to describe, although words cannot do it justice.

I acknowledge Instituto de Telecomunicações (IT) for hosting and supporting my research, and Fundação para a Ciência e Tecnologia (FCT) for funding my PhD under grant SFRH/BD/97204/2013 and for the project UID/EEA/50008/2013 (pluriannual funding in the scope of the LETSREAD project at Instituto de Telecomunicações). I would like to acknowledge the support given by Microsoft to this work.

Abstract

The work detailed in this thesis proposes solutions to automatically evaluate the reading ability of children, targeted towards European Portuguese. Contributions were made towards the state of the art of reading assessment by using sentences and pseudowords, proposing sentence utterance segmentation strategies that consider disfluencies and by proposing multiple features both for mispronunciation classification and overall reading performance estimation.

Knowing how to read is one of the most important markers of a child’s cognitive development. Teachers usually have to expend a large effort to properly evaluate a child's reading aloud performance on a 1-on-1 basis, manually taking notes for accuracy and time. A tool that records and automatically analyzes reading tasks could be an important complement for reading evaluation. The objectives of this work were to develop methods that support an automatic reading evaluation of children 6 to 10 years old. Providing an overall reading aloud level score can be useful to quickly get an appreciation of a child’s ability and follow their evolution along time and to combine the information of several metrics that teachers take into account when evaluating a child.

A large European Portuguese database of children reading aloud was collected to have sufficient data to train acoustic models and to have a large number of examples of reading disfluencies. The reading tasks presented to children were carefully designed by selecting appropriate sentences and generating pseudowords, distributed throughout tasks according to a difficulty metric. Several types of disfluencies were identified, with the most common being mispronunciations, false-starts, repetitions and intra-word pauses. Consequently, these were the ones targeted for automatic detection.

Several strategies were developed to automatically detect reading disfluencies and get automatic annotation of utterances. All followed the same two-step basis: segmentation that detects extra content, and mispronunciation classification. First, segmentation is achieved by constrained decoding lattices based on the ideal pronunciation of the prompt text but allowing freedom of repetition and syllable-based false starts. The best approach uses syllables as units and allows optional silence between each syllable to address the problem of intra-word pauses. Decoding an utterance results in word candidate segments that will be classified as correctly pronounced or not. The best performing feature to classify mispronunciations was a log-likelihood ratio between the ideal pronunciation and a free-phone-loop filler model, done in a word-spotting manner. Additional features of likelihoods of individual phonemes and Levenshtein distances between correct pronunciation and recognized phonemes are combined in multi-feature models.

Elementary school teachers were asked to rate the overall reading level score of children as 0-5, resulting in a ground truth of reading score. Regression models to estimate these scores were trained based on performance features extracted from the automatic annotation, with separate features for the reading of sentences or pseudowords. Gaussian process regression models achieved the best results from automatic annotation, with results closely approaching the use of features extracted from manual annotation.

Two applications of the developed work were built: a demo and a prototype website. The demo application showcases the methodology applied to the children of the collected dataset. The prototype website is a platform for teachers where they can assign reading tasks to several students, tell children to read tasks using a microphone, and analyze the automatically given performance score and utterance annotations.

Keywords: Automatic speech recognition, child speech, disfluency detection, reading aloud performance

Resumo

O trabalho detalhado nesta tese propõe soluções para automaticamente avaliar a capacidade de leitura de crianças, tendo como alvo o Português europeu. Foram feitas contribuições para o estado da arte de avaliação de leitura ao usar frases e pseudopalavras, ao propor estratégias de segmentação de locuções de frases que consideram disfluências e ao propor vários parâmetros quer para classificação de pronunciações incorretas como para estimação de capacidade de leitura geral.

Saber ler é um dos mais importantes marcadores do desenvolvimento cognitivo de uma criança. Os professores têm habitualmente de despender um grande esforço a avaliar decentemente a capacidade de leitura em voz alta de uma criança, um a um, manualmente tirando notas de exatidão e tempo. Uma ferramenta que grave e automaticamente analise tarefas de leitura poderá ser um importante complemento à avaliação de leitura. Os objetivos deste trabalho foram desenvolver métodos que suportem uma avaliação automática da leitura de crianças de 6 a 10 anos de idade. Fornecer um valor de nível geral de leitura em voz alta pode ser útil para rapidamente obter uma apreciação da capacidade de uma criança e seguir a sua evolução ao longo do tempo e para combinar a informação de várias métricas que os professores têm em conta quando avaliam uma criança.

Uma grande base de dados de Português europeu de crianças a ler em voz alta foi adquirida para ter dados suficientes para treino de modelos acústicos e para ter um largo número de exemplos de disfluências da leitura. As tarefas de leitura apresentadas às crianças foram cuidadosamente construídas seleccionando frases apropriadas e gerando pseudopalavras, distribuídas ao longo das tarefas de acordo com uma métricas de dificuldade. Vários tipos de disfluências foram identificados, com os mais comuns sendo pronunciações incorretas, pré-correções, repetições e pausas intra-palavra. Consequentemente, estes foram o alvo de detecção automática.

Foram desenvolvidas várias estratégias para automaticamente detetar disfluências da leitura e obter anotação automática de locuções. Todas seguiram a mesma base de duas fases: segmentação que deteta conteúdo extra, e classificação de pronunciações incorretas. Primeiro, segmentação é conseguida por gramáticas de descodificação restritas baseadas na pronunciação ideal do texto mas permitindo liberdade de repetição e pré-correções baseadas em sílabas. O melhor método usa sílabas como unidades e permite silêncio opcional entre cada sílaba para responder ao problema de pausas intra-palavra. Descodificar uma locução resulta em segmentos candidatos de palavra que serão classificados como correta ou incorretamente pronunciados. O melhor parâmetro para classificar pronunciações incorretas foi uma razão de verosimilhança logarítmica entre a pronunciação ideal e um modelo de enchimento com todos os fones em paralelo, feita de uma forma semelhante a deteção de palavras. Parâmetros adicionais de verosimilhanças individuais de fonemas e de distâncias de Levenshtein entre pronunciação correta e fonemas reconhecidos foram combinados em modelos de múltiplos parâmetros.

Professores do 1º ciclo do ensino básico foram convidados a avaliar o nível geral de leitura de crianças em 0-5, resultando em valores de referência. Modelos de regressão que estimem estes valores foram treinados baseados em parâmetros extraídos da anotação automática, com parâmetros diferentes para a leitura de frases e de pseudopalavras. Modelos de regressão de processos Gaussianos obtiveram os melhores resultados com anotação automática, resultados que se aproximaram do uso de parâmetros extraídos da anotação manual.

Duas aplicações do trabalho desenvolvido foram construídas: uma *demo* e um *website* protótipo. A *demo* apresenta a metodologia desenvolvida aplicada às crianças da base de dados adquirida. O *website* protótipo é uma plataforma para professores onde estes podem atribuir tarefas de leitura a vários alunos, dizer às crianças para lerem tarefas usando um microfone, e analisar o nível de leitura automático e a anotação automática.

Palavras-chave: reconhecimento automático de fala, fala de crianças, deteção de disfluências, capacidade de leitura em voz alta

Contents

List of figures

List of tables

Acronyms

ASR	Automatic Speech Recognition
CALL	Computer-assisted language learning
CG	Curricular goals
CV	Cross-validation
CWPM	Correct words per minute
DET	Detection error tradeoff
DTW	Dynamic time warping
EP	European Portuguese
FST	Finite state transducer
GMM	Gaussian mixture models
GOP	Goodness of pronunciation
GPR	Gaussian process regression
HMM	Hidden Markov models
LASSO	Least absolute shrinkage and selection operator
LLR	Log-likelihood ratio
LR	Linear regression
MAP	Maximum a posteriori
MFCC	Mel-frequency cepstral coefficients
NN	Neural network
OLD20	Orthographic Levenshtein distance 20
PCA	Principal component analysis
PHO	Small mispronunciation events (one phonetic change)
PL	Phonetic lattice
PRE	False start events (or pre-corrections)
QbE-STD	Query by example spoken term detection
QUESST	Query by Example Search on Speech Task

REP	Repetition events
RMSE	Root mean squared error
SSE	Sum of squared errors
STD	Spoken term detection
SUB	Substitution events (mispronunciation)
SVM	Support vector machine
WER	Word error rate

Chapter 1

Introduction

With current technological advancements, computational resources can play a part in most aspects of a child's education and their use is foreseen to increase in the near future. Knowing how to read is an important skill for a child's development and one of the markers of their cognitive capabilities. There are two main avenues where technology can contribute to improve reading aloud ability of children: automatic reading tutors and automatic evaluation. Reading tutors are usually attractive computational applications where the effort goes to following children's reading, giving them feedback and encouraging them to improve their reading. On the other hand, automatic evaluation of a child's reading performance can be an important tool to complement the current method of manual evaluations of overall reading ability in schools. Usually, teachers or tutors need to make the effort of providing a level-appropriate reading task to the child, manually take notes for time and accuracy, and calculate a metric such as correct words per minute. This 1-on-1 procedure can be very time-consuming, especially if additional performance metrics are desired. Also, manual evaluations are not consistently equal and depend on evaluator bias and experience. An automatic system that can perform these steps accurately would be a great complement to the usual methods and an indispensable tool for teachers that may have classes with up to 30 children. It could also lead to more frequent assessments of a child throughout the school year, and a higher-quality accompaniment of their education. Providing an overall performance score, as opposed to specific metrics and subjective parameters, can give a clear overview of a child's status and can also be beneficial for the analysis of a child's progress over time.

1.1 Motivation and objectives

As a PhD topic and having a final application in mind, a decision was made to focus on reading aloud evaluation. Both automatic reading tutors and automatic reading evaluation need to use automatic speech recognition and speech processing technologies, which are the main subjects of this work. The methodology for tutoring and evaluation are similar, because both need to analyze how well a child is reading by detecting reading mistakes. This means that the work developed here could be used to build reading tutors as well. Another possible use of the methodology of reading aloud assessment is to detect reading disorders and find specific problems, which is also not targeted here. These were the initial objectives/tasks proposed:

1. Data: Collect and annotate a database of reading aloud tasks from 1st grade children, rich in disfluencies, for training acoustic models and developing disfluency[1] detection methods;
2. Models: From a set of the collected data, train robust acoustic phoneme models, tailored for children. These are required to decode the audio signal of further utterances of children for disfluency detection;
3. Disfluencies: Develop methods to automatically detect all reading disfluencies (mispronunciations, repetitions, corrections, insertions, intra word pauses, extensions) in children's reading tasks;
4. Reading Ability Index: develop an algorithm/formula to attribute an overall score for a child from several utterances of reading aloud, based on the automatic detection of problems. The score or index should be well correlated with the opinion of teachers or experts.

[1] In this document, the term 'disfluencies' is used to represent all reading miscues and errors and may represent events of extra content as well as mispronunciations.

The objectives proposed for this PhD are closely related to the LetsRead project[2]. The *Instituto de Telecomunicações* internal project, with the collaboration of Microsoft, officially ran from 2014 to 2016, but developments continued, specifically on improvements to automatic detection of disfluencies and the development of a full web application that teachers can use, in addition to the project's demo application.

One of the main motivations for this work was that there were no computer assisted applications for European Portuguese (EP) that automatically evaluate the reading aloud performance of children. Even for other languages, this automatic evaluation is a developing topic. To provide an overall reading ability score, the focus was always on reading of isolated words (Black et al., 2011; Duchateau et al., 2007) and mainly using reading speed and number of correctly read words to estimate the score. Using and analyzing sentences and pseudowords (non-existing/non-sense words) for overall performance scoring is one of the contributions of this work and it is expected that, by working with sentences as well as pseudowords, a better understanding of a child's reading ability can be achieved. New methods to automatically detect disfluencies are developed and regression models are explored to provide performance scores based on multiple sources of information that can be the ones that teachers consider to evaluate children. The work described in this document can be summarized to these topics: design, collection and analysis of a children reading database; automatic detection of reading disfluencies; automatic reading performance evaluation; final live application of the developed methods.

There were several challenges to be tackled in this work. One is the lack of speech data of European Portuguese children reading, which led to the need to collect a new database of reading tasks. Another issue is that the speech of children presents significant differences to the speech of adults for whom speech recognition and speech technologies are relatively mature. Children can also make several distinct types of reading errors and disfluencies, which need to be addressed in an automatic fashion.

Falling under the interest of the thesis work, several efforts were carried out on the topic of Query-by-Example spoken term detection (QbE-STD), from the onset of the PhD. The

[2] The LetsRead project: http://lsi.co.it.pt/spl/projects_letsread.html

objective of QbE-STD is to match the content of a spoken query (usually one or a few words) to a large number of audio files from speech databases, identifying where that query appears (Anguera et al., 2014; Metze et al., 2014; Szoke et al., 2015; Tejedor et al., 2016, 2013). The challenge of spoken term detection 'by example' means that there is no prior textual knowledge of what is being said in the query. The main objective of exploring QbE-STD was that it could be useful to identify where correctly pronounced words are in an utterance. Adding the extra challenging aspect of language independence, the usual strategy stems from using phonetic recognizers of different languages and using dynamic time warping (DTW) to match templates, without any textual inference of what is being said. Search 'by example' would ultimately not become an avenue to pursuit for the problem at hand of analyzing children's reading attempts, since the focus became of detecting deviations to the intended reading, given by a prompt, and not of matching two spoken examples. However, the work developed for QbE-STD was of utmost relevance to acquire knowledge about training neural networks for phonetic recognition, working with posterior probabilities of phones and dealing with event detection issues – all pertinently applied in the detection of disfluencies in read utterances. The subject of QbE-STD is not described in this thesis, keeping the theme focused on children evaluation. Nevertheless, the contributions made about the topic will be summarily reported.

The reader will find that this document is not focused on introducing speech recognition basics and topics such as deep neural networks. When describing disfluency detection and automatic transcription (chapter 3), it is assumed that the reader has some familiarity with automatic speech recognition. While not making developments towards the state-of-the-art of acoustic modelling, the focus is on new strategies for automatic transcription and using multiple features for mispronunciation classification and performance evaluation, with several machine learning approaches. The challenges of evaluating the reading ability of children and characteristics of their speech are discussed in this introductory section.

1.2 Reading ability of children

In the process of learning how to read, children can face phonological, phonic or rhythmical difficulties in reading aloud, reflecting different levels of fluency (Lopes et al., 2014;

National Reading Panel, 2000). Oral reading fluency depends on speed, accuracy, consistency of pace and expressiveness (Fuchs et al., 2001; National Reading Panel, 2000). It should be emphasized that this work targets oral reading fluency evaluation, and no effort is made to measure comprehension of what is being read. Nevertheless, there is evidence that oral reading fluency is an indicator of overall reading competence (Fuchs et al., 2001).

The topic of language learning is most closely related to the field of educational sciences. Language learning is a complex cognitive process and there are several specific components that a child has to acquire to be able to process and read written text accurately: phonologic conscience, rules of morphology, syntax, semantics and an extensive vocabulary. Although the first grade of elementary school (children are 5-6 years old at the start of the school year) is when the learning of reading usually has its formal beginning, this often starts earlier on some children, depending on their upbringing or other factors. However, teachers need to expect that children will start to learn how to read at this stage. There are also distinct internal processes for reading, or, more exactly, evolving processes for reading. For example, some children at the end of the 1st grade still read syllable by syllable, and may only acquire the ability to visually recognize full words later on. Children also acquire an extensive vocabulary throughout the years, and the visual recognition of known words is correlated to how fast or how well a word is pronounced. This lexical acquisition has been studied for European Portuguese by the P-PAL project[3], and a lexical database with grade-level word statistics has been collected - ESCOLEX (Soares et al., 2014). This gives insight to the age-of-acquisition of some words, which may be related to the difficulty of reading such words.

By the end of the school year, children are expected to have a certain level of knowledge and skills that reflect what they have learned. The Portuguese government has defined some of these curricular goals (*metas curriculares*) that include qualitative and quantitative objectives for reading aloud (Buescu et al., 2015). Some of these objectives are broad such as “reading with sufficiently correct articulation and intonation”. However, there are some goals that define target values of correct words per minute of different reading tasks,

[3] http://p-pal.di.uminho.pt/about/project

summarized in Table 1.1: text/story words, list of random words from a text and list of pseudowords.

Table 1.1: Correct words per minute targets from curricular goals (*metas curriculares*) for three reading aloud tasks.

Grade	Text	Words	Pseudowords
1st	55	40	25
2nd	90	65	35
3rd	110	80	-
4th	125	95	-

Pseudowords are sequences of characters or syllables that make up "legal" and fully pronounceable words that do not belong to the lexicon and thus have no meaning. They should respect the language's phonotactic constrains that describe how the phonemes of a language can organize to form syllables and words (Mateus and Pardal, 2000). The reading of pseudowords is useful to assess if a child is aware of and can correctly apply the language's morphological and phonemic rules. Notice from Table 1.1 that for the 3rd and 4th grades, no goals for a pseudoword reading task are defined. This could be due to the expectation that the fundamental morphological and phonetic rules should be internalized at this stage, although the analysis of reading performance of pseudowords for these grades could be relevant to assess if further progression is made.

The defined curricular goals can be a starting point to appraise a child's reading ability. However, the Portuguese curricular goals are a controversial topic[4], and there should be other ways to qualify reading performance. One possibility is to gather the opinion of experts and teachers, asking them to quantitatively rate the reading aloud performance of children (by listening to recordings of reading tasks). Their subjective opinion will be based on the several aspects of reading (speed, fluency, number of mispronunciations, etc.), and if these parameters can be quantified, there can be a correlation between the human score and a combination of the parameters. This is the premise of building an overall reading ability

[4] News article in Portuguese: http://www.rtp.pt/noticias/pais/metas-curriculares-do-1-ciclo-sao-atrocidade-cometida-contra-as-criancas_n835820

score that is well correlated with the opinion of expert evaluators (Black et al., 2011; Duchateau et al., 2007).

The use of automatic speech recognition technologies to analyze reading performance gains prominence as an alternative to any kind of manual or 1-on-1 evaluation. The automatic evaluation of literacy or reading ability (not necessarily of children) is always related to detecting correctly read words, or optionally detecting what kind of mistakes are made. Additionally, there are several systems oriented to improve the literacy of an individual (Abdou et al., 2006; Probst et al. 2002), ideally denoting and warning about reading errors that occur. Computer-Assisted Language Learning (CALL) is the area of research that focuses on this subject, allowing a self-practice or an oriented training of the language. These systems are most often created for foreign language learning (Abdou et al., 2006; Cincarek et al., 2009; Franco et al., 2010), therefore, targeted to adults or young adults for whom automatic speech recognition and speech technologies are significantly mature. Nevertheless, for children, there are also applications that deal with the improvement of reading aloud performance – reading tutors. Some projects aimed to create an automatic reading tutor that follows and analyzes a child's reading, such as LISTEN (Mostow et al., 1994), Tball (Black et al., 2007), SPACE (Duchateau et al., 2009) and FLORA (Bolaños et al., 2011). Most of these applications are helpers, e.g., highlighting words in a sentence as they are correctly pronounced.

The final goal of the thesis work is to have an application that can evaluate children's reading ability automatically, and not necessarily provide feedback to them, but to teachers and tutors. It should be mentioned that there is another important application of speech technologies for children in computer-aided speech therapy or diagnosis for impaired individuals (Maier et al., 2009; Saz et al., 2009). However, in this work, no focus will be made for children with specific reading disabilities, so only healthy unimpaired children will be considered. For EP, there are still no robust CALL applications. Nevertheless, some systems have been developed for EP that pursue the analysis of children's reading in an interactive fashion, such as REAP.PT (Silva et al., 2012). However, it deals with issues such as readability of texts and is geared towards comprehension and not reading aloud ability.

As for the automatic detection of correct words and of disfluencies in children's reading, most works are based on having decoding networks (or lattices) that allow the fully correct word or sequence of words to be detected, concurrently with alternative events that should be reading miscues. Other connections in these lattices can allow, e.g., repetitions of words. The state-of-the-art on this topic will be described in chapter 3. Most works focus on individual word reading, and the ones that aim to provide an overall reading ability score use individual word reading exclusively, whereas this PhD's objective is to use sentences and pseudowords for that goal. Results of disfluency detection tasks in literature show that there is still a significant room for improvement in this topic.

1.3 Child speech

The speech of children, due to their smaller vocal tracts, presents certain acoustic characteristics that are significantly different from adult speech such as higher fundamental and formant frequencies, formant frequency variability and vowel duration variability (Ferreira, 2007; Hämäläinen et al., 2014b; Lee et al., 1999). Acoustic models trained with adult speech may have a diminished performance in automatic speech recognition (ASR) applications for child speech, and special care is needed to adapt or create models that target children (Hämäläinen et al., 2014a; Potamianos and Narayanan, 2003). Several methods can be found in the literature to deal with this issue: retraining adult models using a small amount of children data; frequency warping with vocal tract length normalization, and other speaker adaptation methods (Gray et al., 2014; Liao et al., 2015). However, it has been suggested that using large amounts of children data may be preferable than applying adaptation techniques (Liao et al., 2015).

When considering reading aloud by children, the deviations to an appropriate reading include reading syllable by syllable, false starts followed by self-corrections, severe mispronunciations of words, and others. The wide range of possible problematic events presents a substantial challenge for computational systems that aim to detect these problems automatically. Having data of European Portuguese children reading aloud seems vital to allow both the creation of robust children's speech acoustic models for EP and the identification of the common disfluencies that children commit while reading, pursuing reading ability evaluation. Although there are some databases for EP such as Speecon with rich sentences[5]; ChildCAST (Lopes et al., 2012) with picture naming; the Contents for Next Generation (CNG) Corpus targeting interactive games (Hämäläinen et al., 2013) and Santos et al. (2014) with child-adult interaction, these do not present the necessary disfluent reading speech needed for the goals of this thesis. Therefore, a database was collected of 6-10 year old children reading aloud sentences and pseudowords, in alignment with the goals of the LetsRead project. Further discussion about child reading and state-of-the-art of automatically detecting reading mistakes and evaluating reading level will be included throughout this document's chapters.

1.4 Chapter overview

This document is structured in the following fashion:

Chapter 2 - Reading Tasks. This chapter revolves around the design and collection of a new database of children reading aloud. It will describe the effort needed to design reading tasks of children, specifically of sentences and pseudowords, along with a description of the data collection process and of the data itself. The performed manual transcription is described and all the types of reading disfluencies found are enumerated. Finally, an analysis of reading performance metrics for the database's children is done.

Chapter 3 - Automatic detection of disfluencies. This chapter will describe the systems developed for an automatic transcription of utterances of children reading where several types of disfluencies are targeted for detection (intra-word pauses, repetitions, false-starts and mispronunciations). The problem was tackled as a two-step process: first, a segmentation stage provides word candidate segments while detecting repetitions and false-starts, and different decoding approaches are discussed; then, a classification stage decides if a word is mispronounced or not and multiple features were defined and combined for classification. The output of the automatic disfluency detection system allows several

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

reading performance metrics to be extracted for the next topic (correct words per minute, rate of disfluencies, etc.).

Chapter 4 - Automatic Assessment of Reading Level. This chapter will describe the strategy to automatically provide an overall assessment of a child's reading level. A ground truth of reading level of the database's children was collected from elementary school teachers, and machine learning methods are applied on automatically extracted features to closely approach the opinion of teachers. Two live applications of the methods developed to assess reading level will be presented.

Chapter 5 - Conclusions. On the final chapter, final remarks and a perspective of future work on the thesis's topic are given, along with a discussion of the most challenging aspects found for evaluating the reading ability of children.

Figure 1.1 shows the workflow of the steps taken to be able to automatically estimate the reading level of children and on which chapters these steps will be detailed.

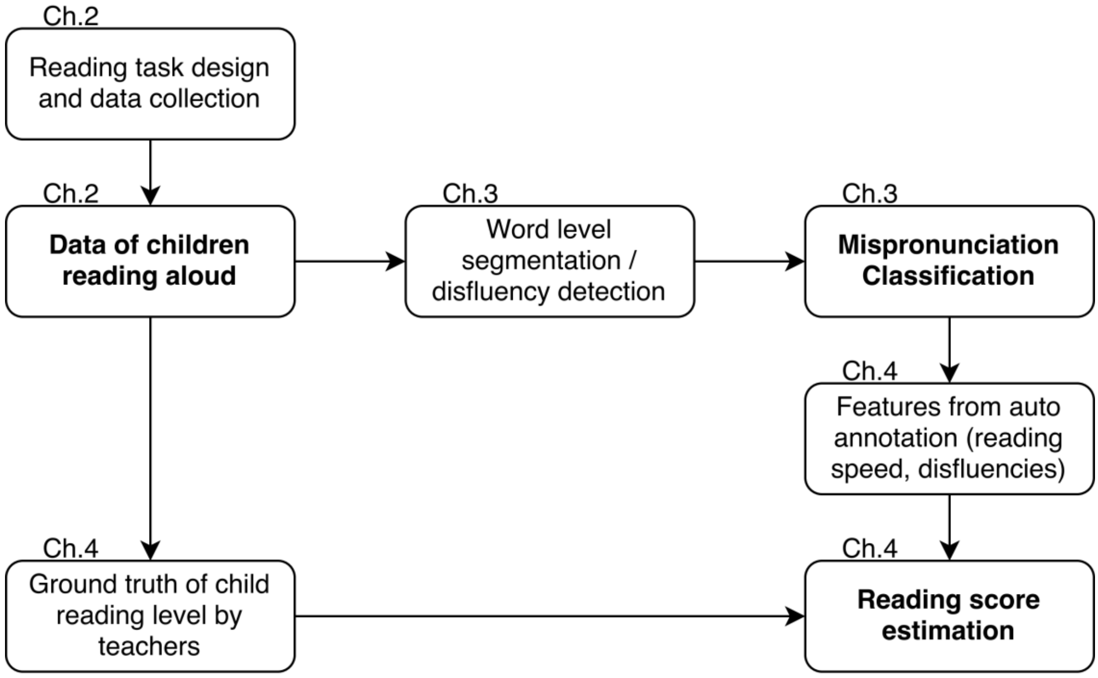

Ch.2
Reading task design and data collection
Ch.2
Data of children reading aloud
Ch.3
Word level segmentation / disfluency detection
Ch.3
Mispronunciation Classification
Ch.4
Features from auto annotation (reading speed, disfluencies)
Ch.4
Ground truth of child reading level by teachers
Ch.4
Reading score estimation

Figure 1.1: Schematic of steps taken to automatically assess child reading level along with corresponding chapters.

1.5 Contributions

The main contributions of this work, reported in 2 journal papers and 8 conference papers, can be defined as:

- The collection of a new European Portuguese speech database of children reading aloud. The collected data of 6-10 years old children reading sentences and pseudowords is full of examples of reading mistakes and disfluencies and the children also vary greatly in reading capability. The audio amounts to 20 hours, more than half was fully manually annotated and the dataset was made available publicly[6]. The database was presented and analyzed in Proença et al. (2015c, 2016a, 2016d).

- Novel strategies to automatically detect extra content in utterances and deal with intra-word pauses. Several decoding grammars were proposed to segment utterances and detect extra events, reported in Proença et al. (2015b, 2015d, 2016b, 2017a). Ultimately a syllable-based decoding grammar had the best performance, reported in Proença et al. (2018).

- New features and feature combination to classify word candidates as mispronounced or not. Several features were proposed based on word likelihood, recognition likelihood and phonetic recognition to improve mispronunciation classification, reported in Proença et al. (2018).

- Overall reading performance estimation based on multiple features from sentence and pseudoword reading. Separate features from sentences and pseudowords based on reading speed, frequency of disfluencies and task difficulty are used in regression models to get closer to the opinion of teachers for a child's reading level. This work was reported in Proença et al. (2017b, 2017c).

- Live demo of the methodology developed applied to children's utterances[7].

[6] The LetsRead Database: http://lsi.co.it.pt/spl/letsreadDB.html

[7] The LetsRead demo: http://hades.co.it.pt:9000/index_en.html – A short video with an overview of the project and of the demo can be found at: http://lsi.co.it.pt/spl/letsread/Letsread_demo.mp4

- Prototype website for teachers to propose reading tasks to children and check the automatic analysis of their reading performance[8].

Furthermore, work developed in the topic of Query-by-example spoken term detection (QbE-STD) resulted in a 2nd place at IberSpeech 2016 ALBAYZIN Search on Speech evaluation[9], 3rd place at MediaEval 2015 query-by-example search on speech task (QUESST) (Szoke et al., 2015) and an honorable mention for “best results with low resources” at MediaEval 2014 QUESST (Anguera et al., 2014). The main contributions to the topic were:

- Novel dynamic time warping (DTW) strategies for inexact spoken queries allowing complex matches with DTW paths with extra content, word reordering and partial words (Proença et al., 2015e, 2014).
- Inclusion of extra sub-systems for fusion based on the average distance matrix of all phonetic recognizers (Proença et al., 2015a; Proença and Perdigão, 2016a).
- Candidate match selection based on document dependent thresholds (Proença and Perdigão, 2016b).
- Recording of European Portuguese queries and data preparation used by the MediaEval 2015 evaluation (Szoke et al., 2015).

[8] LetsRead prototype website: https://letsread.co.it.pt/ – The website is only available in Portuguese.

[9] ALBAYZIN 2016 Search on Speech: https://iberspeech2016.inesc-id.pt/index.php/albayzin-evaluation/

Chapter 2

Reading Tasks

As discussed in the introductory chapter, there was a need to acquire a new database of European Portuguese children reading aloud in order to achieve the goals of this work. The main specific requirements of data collection for the study of reading performance were:

- Reading without a chance to practice before-hand. If a child can see the prompt before the start of recording, reading performance may be higher and there can be no account of reaction time. This is in contrast to what is usually done in speech data collections where it is ideal to have a correct reading of prompts, so that it is easier to build acoustic models. The reaction time of starting to read a presented sentence or word can be indicative of reading level and have impact on reading speed.
- Have varied material for the text of prompts, with diverse lengths and difficulty.
- Have quality recordings with low noise and reverberation. It is not the objective to solve acoustic quality issues.
- Get recordings from a large number of subjects at an elementary school level (1st to 4th grades), ideally with similar numbers per gender and grade.

This would allow for the collection of a large number of examples of reading mistakes and other disfluencies, examples needed to train automatic disfluency detection systems (which will be vital to automatic overall reading performance evaluation). Before data collection, the main issue was then to propose reading tasks to be read. These tasks should be carefully designed and should be well balanced so that if a child reads one task, their performance should be similar to reading a separate task.

As previously stated, the Portuguese government has defined a set of Curricular Goals (CG) with qualitative and quantitative objectives per grade for reading aloud (Buescu et al., 2015). Some of these objectives include target reading speed of words per minute on short texts, individual words and pseudowords reading tasks. With the analysis of curricular goals in mind, utterances consisting of read sentences and pseudowords were the goal of data to be collected. It was decided not to include reading of isolated words, as the required time for a session could become too long and the child's performance is likely to decrease with extended sessions. Additionally, similar works for reading level estimation use individual word tasks (Black et al., 2011; Duchateau et al., 2007) and it was expected that a more thorough analysis of reading ability could be done by using sentences and pseudowords[10].

The pseudoword reading task was included as it may provide a different and objective analysis of phonetic awareness and letter-to-sound rules independent of word familiarity and context. With sentences, plenty of reading disfluencies can be collected from which the overall reading performance of a child can be evaluated. For the LetsRead database collection, each child was presented with a reading task that asked them to read aloud twenty sentences and ten pseudowords. Initially, forty reading tasks were established (rotating 10 per grade) to balance repetition and diversity of the data. At a later stage of data collection, these were shortened to 5 tasks per grade, to reinforce repetition of contents. The vocabulary of the set of sentences and pseudowords comprises a total of 3324 word types. Examples of the defined reading tasks for each grade can be found in Annex I. The distribution of the material for the different grades was made according to a difficulty metric.

2.1 Difficulty

The difficulty of reading a word or sentence is an important parameter to take into account while designing reading tasks. The expectation of what a first grader can read is different

[10] In hindsight, it would have been interesting to also collect samples of reading individual real words to be able to directly compare results to other works. However, the justification of needing short collection times remains valid.

than for a fourth grader and, usually, they are evaluated on different texts. A metric of difficulty was defined and attributed to words and sentences to make reading tasks averagely harder as the grade increases and to make the prompts of a reading task diverse in difficulty.

In literature, difficulty of reading often appears associated with silent reading and not necessary oral pronunciation. Popular metrics such as readability define how hard or advanced a text is to understand (Mc Laughlin, 1969; Silva et al., 2012), but there is no consideration for how hard it will be to read aloud for a child. For oral production, there are studies that consider pronounceability or phonetic complexity but are usually targeted towards speech disorders such as stuttering and dyslexia (Bose et al., 2011). The identified issues that may have impact on the reading aloud difficulty of words are:

- Pronounceability – articulatory aspects of speech production;
- Phonetic complexity;
- Number of syllables and syllable encounters;
- Word familiarity – age of acquisition, frequency of the word in the language, or similarity to frequent words;

At the start of the study, for defining reading tasks, only phonetic complexity and syllable based rules are considered. A similar approach to Jakielski (1998) was taken, who defined an index of phonetic complexity attributing difficulty points to words based on factors such as consonant types, word length, consonant clusters and consonant-vowel structure. Although it would be ideal to also relate a word's difficulty to its age-of-acquisition or familiarity, not all words of the proposed reading tasks were present in available lexical databases such as ESCOLEX (Soares et al., 2014), and considering such features was not directly feasible. For example, 'Daniel' is a common given name and it is badly represented for its non-existing frequency in ESCOLEX. For later tasks, additional considerations will be made for difficulty, such as orthographic Levenshtein distance 20 (OLD20) (Yarkoni et al., 2008), which defines the distance of a word to its closest 20 orthographic neighbors, which is indicative of similarity to other words and may relate to a perception of familiarity.

It was decided to build a parameter of difficulty where sentences are evaluated in terms of phonetic complexity and variety. All words were split into their syllables and a difficulty level was assigned to each syllable, summing the scores given by the structural, grapheme and syllable encounter rules defined in Table 2.1.

Table 2.1: Rules for syllable difficulty score (C – consonant, V – vowel, N – nasal vowel)

Syllable rules	**Added score**	**Syllable example**	**Word example**
V	0	a, o	aba, ovo
CV	0	ca, po	cama, sapo
VC	1	al, er	altar, ermida
VV	1	ou, ei	outro, eito
CN	2	man, cam	manto, campa
CVC for a plural	0	cos, sas	barcos, casas
CVC	1	ter, bor	meter, sabor
CVV or VVC	1	dou, ais	doutor, sociais
VCC or CCV	2	plu	plural
CVN	2	quan, guin	quando, guindaste
CCN	3	plan, bran	planta, lembrança
CNC	4	vens	vens
5 graphemes	5	trans	transporte
C(r|l)	1	plu, flo, dra	plural, florescer, quadra
‘ch’ or ’qu’ or ’gu’	1	cha, quin, gus	chama, quintal, gustar
‘nh’,’lh’	2	nhe, lho	amanhecer, olho
‘x’ on a 1st syllable	1	xa	xadrez
‘x’ on a syllable’s end	3	fax, lix	fax, felix
‘x’ on other cases	5	xe, xa, xo	trouxe, exame, baixo, fixo
‘s’ on a syllable’s start, preceded and followed by a vowel	1	sa	casa
‘r’ on a syllable’s start, preceded by ‘n’ or ‘l’	3	ra, ro	honra, melro
‘c’ or ‘g’ on a syllable’s start, followed by ‘e’ or ‘i’	1	cel, ci, gi, ge	pincel, cima, giro, age
‘em’ or ‘am’ at the end of a word not preceded by ‘e’	1	lem, cam	falem, socam
‘em’ preceded by ‘e’, or ‘êm’ at the end of a word	3	em	vêm, veem
C on a syllable’s start, preceded by C	1	pa, ma	caspa, arma
V on a syllable’s start, preceded by V	1	ei	candeeiro

For each word, the difficulty scores given to its syllables are summed, as well as a ⅓ value per syllable so that word length is also accounted for (a word with 3 syllables would have an added difficulty of 1). Further on, the difficulty metric will be used as a feature for mispronunciation classification, and it will be additionally considered without the length rule, meaning that there can be words of null difficulty (mostly CV-CV words). As an example of syllable difficulty without accounting for length, for the sentence "*O Jorge desenha uma figurita*" [u ʒˈɔɾʒə dəzˈɐɲɐ ˈumɐ figuɾˈitɐ][11]:

O	Jor	·ge	de	·se	·nha	u	·ma	fi	·gu	·ri	·ta
0	1	2	0	1	4	0	0	0	1	0	0

The sentence's difficulty would be 9 without accounting for length, or 13 by accounting for its 12 syllables.

After data collection and transcription, an attempt was made to alternatively compute word difficulty based on the frequency of disfluent events for each word. It is expected that words for which more problems were found in attempts of reading them aloud should be of higher difficulty. In reality, due to the high number of different material in the defined reading tasks, there were many words that are apparently difficult but did not present any disfluencies. This fact will be discussed further on, and it will be apparent that there is some correlation of the proposed difficulty metric to the amount of reading disfluencies.

2.2 Sentences

A large set of sentences was extracted from children's tales and school books and exercises of the level of the target group (6-10 years old, 1st-4th grades). Additionally, a small number of sentences from more advanced contents were collected, so that harder content would be included, probably with unfamiliar words. The selected sentences were mostly short, with a maximum length of 30 words and a mean and standard deviation of 11.1±5.8 words. Figure 2.1 shows a histogram of sentence length given in number words. Twenty sentences were included in each reading task (for a recording session with one child). This number was selected for an estimated time per recording session of 10 minutes.

[11] In this document, stress is marked before the stressed vowel and not the syllable.

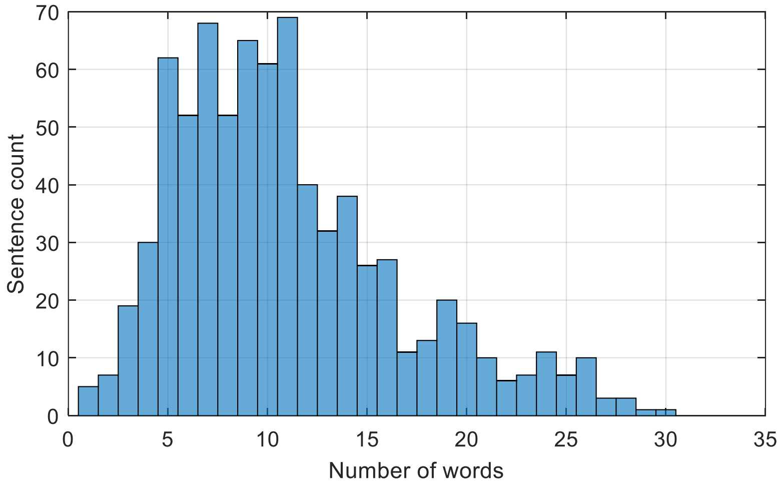


Figure 2.1: Histogram count of considered sentences per number of words.

Since sentences represent the bulk of material collected, the first concern for distributing sentences along the reading tasks of each grade was to maintain a good representation of all phones, so that acoustic models of good quality could be built from the data, similarly to Mendonça et al. (2014). The idea was to not leave out any phone and keep high frequencies for the least represented ones in the gathered material. For example, for the sentences gathered for a certain grade, the amount of phones of their phonetic transcription is analyzed and a subset of 100 sentences is selected (for 5 reading tasks with 20 sentences each), starting with the ones that contain the least represented phones. The considered phonetic alphabet is described in Annex II. An example of the number of phones for 200 sentences for the second grade is shown in Figure 2.2, along with the number of phones for the 100 sentences selected, where the priority given to less frequent phones is apparent. Stressed and unstressed phones were separated here.

The other main concerns in building appropriate reading tasks were to maintain the same average difficulty within a grade (with a rising average difficulty from 1st to 4th grades) and to have sentences of varying difficulty within a task (resulting in overlapping distributions of difficulty for different grades). Furthermore, it is necessary to elicit all types of reading disfluencies as training samples, so the difficulty cannot be too low, although a balance must exist so as not to make the tasks unduly hard.

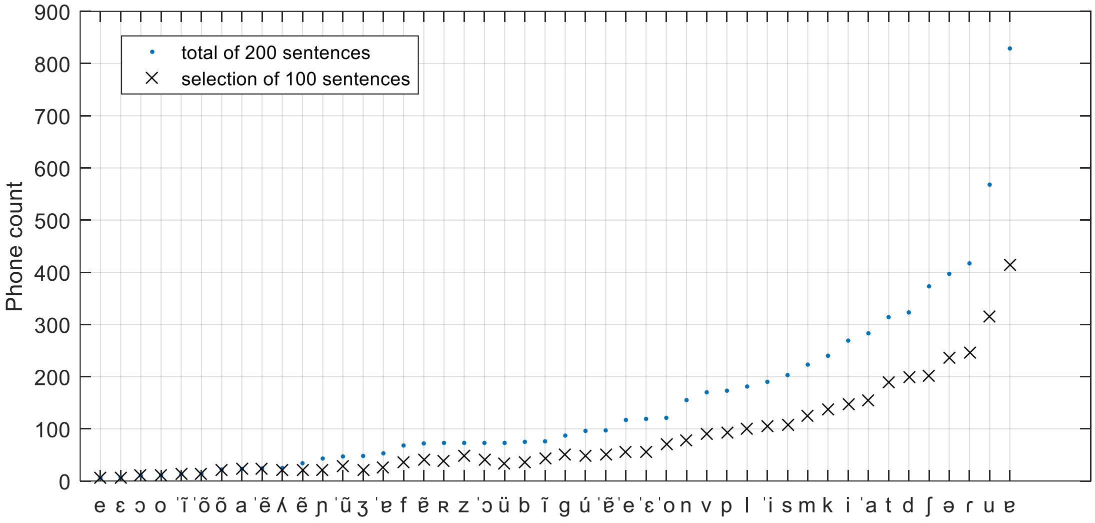


Figure 2.2: Phonetic distribution of 200 sentences and of the 100 selected ones for second grade reading tasks.

2.3 Pseudowords

Pseudowords such as <traba> [tɾˈabɐ], <impemba> [ĩpˈẽbɐ] or <culenes> [kulˈɛnəʃ], represent non-existing or nonsense words. They can be used to evaluate morphological and phonemic awareness of the child and word familiarity does not play a part while attempting to read these words. It is expected that the information extracted from the performance of reading pseudowords may be complementary to the performance of reading sentences, even if they appear correlated.

A new method for the creation of pseudowords was developed. Existing tools such as Wuggy (Keuleers and Brysbaert, 2010) take as input existing words and output pseudowords that differ in one or two syllables to the original words. This creates pronounceable words that are similar to existing words (such as <sapado> [sɐpˈadu] from <sapato> [sɐpˈatu]). The proposed method creates pseudowords without the starting point of valid words while maintaining full pronounceability. It should create non-existing words and the difficulty of reading them is expected to be higher than familiar words. The aim was to create pseudowords of two, three and four syllables. First, the most frequent syllables in each position for words with those number of syllables were extracted from a large lexicon of European Portuguese, CETEMPúblico (Rocha and Santos, 2000). A total of 1754 unique syllables were identified at a grapheme level. Then, words of two or more syllables are created randomly from a set of the most frequent syllables, following some restrictions. For example, to create a 3-syllable pseudoword:

- For the first syllable, only syllables that exist in an initial position are allowed;
- For the second syllable, only syllables that exist in the second-to-last position are allowed;
- For the last syllable, only ones that exist in an ending position are allowed;
- A syllable is randomly selected but with higher probability for higher frequency syllables of the considered position;
- There can only be one syllable with an accented grapheme (e.g.: <ô>, <ã>);
- A syllable cannot begin with the same letter as the final one of the previous syllable;
- If the previous syllable ends with a consonant, the current one cannot start with a vowel;
- The previous syllable must end in a vowel for the current one to start with <ss> or <rr>;
- If the previous syllable ends with <m>, the current one must start with <b> or <p>;
- If the previous syllable ends with <n>, the current one cannot start with <m>, <b> or <p>;
- If the word is accented in the antepenultimate syllable, the last syllable must not end with <i>, <u>, <is>, <us>, <l> or <r>;
- Words that exist in the lexicon (which may be created by chance) are removed.

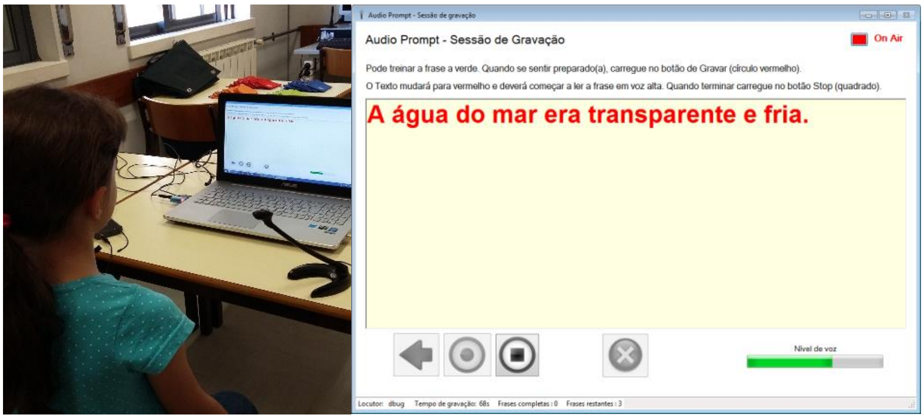
Furthermore, words that presented other syllabic combinations that do not respect pronounceability rules were manually excluded. Pseudowords such as ‘adodo’, ‘catido’ and ‘detate’ are examples of generated words whose syllables are highly frequent and ‘denrambol’, ‘posploei’ and ‘truambro’ are examples of words whose syllables are less frequent.

The difficulty score for a pseudoword is calculated by the previously described method and their distribution along the reading tasks is also similar to sentences, promoting a range of difficulty and rising average difficulty along the grades.

2.4 Data Collection

The corpus of children reading aloud was collected at two private and nine public schools in urban centers and peripheral areas of Portugal's central region with children attending elementary school, aged 6 to 10 years. It was not feasible to collect data outside this region, even if there are certain regional varieties to consider in the country. However, some variations will be allowed as alternative pronunciations during automatic detection of disfluencies.

The end of the school year was targeted (June), to allow first grade children a full year to learn how to read and to better analyze the expected reading level at the end of a grade. A specific application[12] was developed in which the sentences are displayed in a large font size on a computer screen simultaneously with the start of recording. This presentation allows no practice time that would influence performance. A screenshot of the application can be seen in Figure 2.3 as well as an example of the recording environment.


Audio Prompt - Sessão de gravação
Audio Prompt - Sessão de Gravação
🟥 On Air
Nível de voz

Pode treinar a frase a verde. Quando se sentir preparado(a), carregue no botão de Gravar (círculo vermelho).

O Texto mudará para vermelho e deverá começar a ler a frase em voz alta. Quando terminar carregue no botão Stop (quadrado).

A água do mar era transparente e fria.

Locutor: dbug Tempo de gravação: 68s Frases completas : 0 Frases restantes : 3

Figure 2.3: Example of the recording environment (left) and software (right).

[12] Audio Prompt available at http://lsi.co.it.pt/spl/resources.html under "Software (Windows) to record utterances from multiple microphones".

The recordings were performed in school classrooms chosen for their low reverberation and noise acoustics. Children were brought to the recording stage in pairs, to lower intimidation and foster cooperation, and were asked to read aloud a set of 20 sentences and 10 individual pseudowords. A lapel Lavalier microphone (Shure WL93) was used as the main recording device, accompanied by a standard table-top PC microphone as backup (Plantronics Audio 10). The background noise could not always be controlled completely, but was mostly low, with an average signal-to-noise ratio of 30 dB, also because the main recording microphone did a good job of filtering out background noise.

The collected database consists of about 20 hours of 7418 utterances of recorded speech from 284 children, 147 female and 137 male, distributed from the 1st to the 4th grade with 68, 88, 76 and 52 children, respectively. The LetsRead database is available[13] with audio files, meta-data and descriptive documentation. It falls under the Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0) (Figure 2.4).

Figure 2.4: The LetsRead logo (left) and Creative Commons license for the database (right).

A set of 4408 utterances of sentences and pseudowords from 213 children has been fully annotated, totaling 10.5 hours. The children (101 male and 112 female) are distributed along the four grade levels with 47, 52, 51 and 33 speakers (children from the 1st to 4th grades, respectively). This data was manually annotated in terms of correct words, mispronunciations and other disfluency events, with further details in the next subsection. Words and events have time alignments and for events that deviate from a word's pronunciation, a phonetic transcription was included, making the data suitable for training acoustic models. These are the data sets defined for different stages of the study:

- Full set of 10.5 hours: analysis of disfluencies and reading metrics (section 2.6).

[13] The LetsRead database: http://lsi.co.it.pt/spl/letsreadDB.html

- Subset of 9 hours: training set of 173 children used to train acoustic models for word and phone recognizers and to optimize decoders and mispronunciation classifiers (chapter 3).
- Subset of 1.5 hours: test set of 40 children (10 per grade) with different speakers than the 9 hour training set (chapter 3).
- Subset from 150 children of 5 sentences and 5 pseudowords each: reading level assessment (chapter 4).

2.5 Manual Transcription

The annotated speech exhibits a great variety of disfluencies that represent the most common types of errors in reading aloud by children. Based on the HESITA database (Candeias et al. 2013) of the HE[eee]SIT[u]ATION project[14], the rules for the annotation and labelling procedure were defined and the full description of transcription guidelines is available at the database's website[15]. An example of an annotated utterance using the Transcriber software[16] is shown in Figure 2.5.

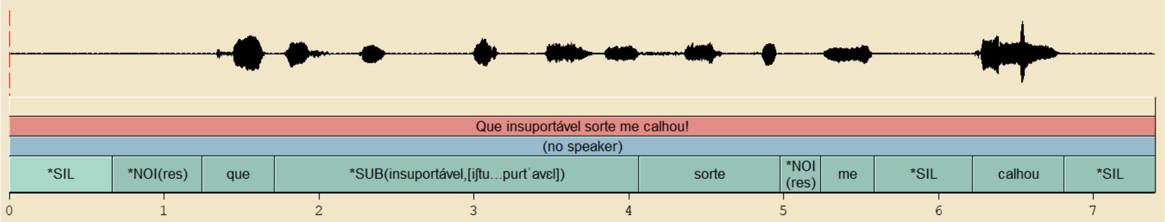

Que insuportável sorte me calhou!
(no speaker)
*SIL
*NOI(res)
que
*SUB(insuportável,[iʃtu...purˈavɛl])
sorte
*NOI (res)
me
*SIL
calhou
*SIL
0
1
2
3
4
5
6
7

Figure 2.5: Example of an annotated utterance of the LetsRead database.

Manual transcription is an arduous process, even more so if a high level of detail is needed. For the LetsRead database, transcription was done in two stages. For the first 5h30 hours, the starting point was as a forced alignment of the word sequence of the utterance's prompt. For the extra 5 hours, a version of the automatic system that detects repetitions,

[14] The HE[eee]SIT[u]ATION project: http://lsi.co.it.pt/spl/hesitation/
[15] Transcription guidelines: http://lsi.co.it.pt/spl/letsreaddb.html under 'Further description'.
[16] Transciber 1.5.1: http://trans.sourceforge.net

false-starts, intra-word pauses and mispronunciations was used to automatically obtain a prototype annotation that still needed to be checked and altered thoroughly.

The annotated speech exhibits a great variety of disfluencies that represent the most common types of errors in reading aloud by children. The several types of disfluency were identified as follows:

- PRE – False starts (or pre-corrections, *reparandums*) that are followed by the attempted correction, where multiple can occur. This follows the theory of there being an interruption point after a *reparandum*, followed by the repair (Moniz et al., 2014; Shriberg, 1994). Example: for prompt ˈ*grande espanto*ˈ [gɾˈɐ̃də iʃpˈɐ̃tu], utterance is ˈ*grande* ***espa*** *espanto*ˈ [gɾˈɐ̃də **ˈiʃpɐ** iʃˈpɐ̃tu].

- REP – Repetition of a word (multiple repetitions may occur). Example: for prompt ˈ*Ele já me deu*ˈ [ˈelə ʒa mə dew], utterance is ˈ*Ele,* ***ele*** *já me deu*ˈ [ˈelə **ˈelə** ʒa mə dew].

- SUB – Substitution or severe mispronunciation of a word. Example: for prompt ˈ*voava em largos círculos*ˈ [vuˈavɐ ẽj̃ lˈaɾguʃ sˈiɾkuluʃ], utterance is ˈ*voava em* ***lares sicos***ˈ [vuˈavɐ ẽj̃ **lˈaɾəʃ sˈikuʃ**].

- PHO – Small mispronunciation of a word, exactly with a change in one phone. Example: for prompt ˈ*A Lena chegou a casa, da escola*ˈ [ɐ lˈenɐ ʃəgˈo ɐ kˈazɐ dɐ iʃkˈɔ.lɐ], utterance is ˈ*A Lena* ***chegou*** *a casa, da* ***escola***ˈ [ɐ lˈenɐ **ʃɐgˈo** ɐ kˈazɐ dɐ **iʃkˈalɐ**].

- INS – An inserted word that is not part of the original sentence. Example: for prompt ˈ*mas também dizem*ˈ [mɐʃ tɐ̃bˈɐ̃j̃ dˈizɐ̃j̃], utterance is ˈ*mas também* ***me*** *dizem*ˈ [mɐʃ tɐ̃bˈɐ̃j̃ **mə** dˈizɐ̃j̃].

- DEL – The word was not pronounced (deletion). Example: for prompt ˈ*onde morava* ***uma*** *velha*ˈ [ˈõdə muɾˈavɐ **ˈumɐ** vˈɛʎɐ], utterance is ˈ*onde morava velha*ˈ [ˈõdə muɾˈavɐ vˈɛʎɐ].

- CUT – The word is cut off, usually in the initial or final syllable, but not corrected later. Example: for prompt ˈ*dá água ao papagaio*ˈ [da ˈagwɐ aw pɐpɐgˈaju], utterance is ˈ*dá água ao **papaga***ˈ [da ˈagwɐ aw **pɐpɐgˈa**].
- EXT – Phonetic extension, marked with the symbol [:]. Example: for prompt ˈ*que queres dizer?*ˈ [kə kˈɛɾəʃ dizˈeɾ], utterance is ˈ***que** queres dizer*ˈ [**kə:** kˈɛɾəʃ dizˈeɾ].
- PAU (…) – Intra-word pause, when a word is pronounced syllable by syllable with intervening silences. The symbol [...] can also appear in other disfluency events denoting a pause. Example: for prompt ˈ*formosa e bonitinha*ˈ [fuɾmˈɔzɐ i bunitˈiɲɐ], utterance is ˈ*formosa e **boni...tinha***ˈ [fuɾmˈɔzɐ i **buni...tˈiɲɐ**].

Silence and non-speech events such as breathing, labial and background noise were also annotated. Extensions and intra-word pauses may occur simultaneously with other disfluencies and are marked with [:] and […] in phonetic transcriptions. Some of the defined tags can overlap in broad definitions, such as SUB, PHO and CUT all being mispronunciations.

2.6 Analysis of Performance

From the manually annotated data, an analysis of certain metrics of children's reading aloud performance can be done. The overall number of disfluencies encountered will be analyzed, along with reading speed metrics compared to curricular goals and pseudoword reaction times and disfluency rates.

2.6.1 Disfluencies

The number of occurrences for each type of disfluency and their percentage over total number of words are presented in Table 2.2 for sentences and Table 2.3 for pseudowords, for each of the four grade levels.

Since events such as PRE and REP can occur multiple times for the same word, the percentage shown in relation to prompt words could be higher than 100%. By analyzing disfluency counts in sentences, some interesting phenomena can be observed, such as 1st-graders being the ones that exhibit more intra-word pauses and extensions (due to slower

Table 2.2: Distribution of disfluency types in sentences with number of events and percentage over prompt word count (total of 37960).

Tags	1st grade	2nd grade	3rd grade	4th grade	Total
PRE	430 (8.3%)	549 (5.6%)	480 (4.0%)	384 (3.5%)	1843 (4.9%)
REP	162 (3.1%)	181 (1.9%)	225 (1.9%)	203 (1.9%)	771 (2.0%)
SUB	236 (4.5%)	339 (3.5%)	427 (3.5%)	224 (2.1%)	1226 (3.2%)
PHO	250 (4.8%)	283 (2.9%)	316 (2.6%)	208 (1.9%)	1057 (2.8%)
CUT	12 (0.2%)	19 (0.2%)	38 (0.3%)	35 (0.3%)	104 (0.3%)
INS	40 (0.8%)	56 (0.6%)	68 (0.6%)	73 (0.7%)	237 (0.6%)
DEL	6 (0.1%)	17 (0.2%)	22 (0.2%)	56 (0.5%)	101 (0.3%)
EXT	228 (4.4%)	154 (1.6%)	117 (1.0%)	68 (0.6%)	567 (1.5%)
PAU	373 (7.2%)	282 (2.9%)	280 (2.3%)	89 (0.8%)	1024 (2.7%)

Table 2.3: Distribution of disfluency types in pseudowords with number of events and percentage over prompt word count (total of 1866).

Tags	1st grade	2nd grade	3rd grade	4th grade	Total
PRE	78 (20.0%)	69 (12.5%)	66 (12.5%)	61 (15.4%)	274 (14.7%)
REP	1 (0.3%)	1 (0.2%)	3 (0.6%)	0 (0.0%)	5 (0.3%)
SUB	79 (20.3%)	74 (13.4%)	67 (12.6%)	34 (8.6%)	254 (13.6%)
PHO	92 (23.6%)	95 (17.2%)	87 (16.4%)	65 (16.5%)	339 (18.2%)
CUT	1 (0.3%)	0 (0.0%)	1 (0.2%)	0 (0.0%)	2 (0.1%)
INS	9 (2.3%)	10 (1.8%)	7 (1.3%)	2 (0.5%)	28 (1.5%)
DEL	3 (0.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (0.2%)
EXT	73 (18.7%)	41 (7.4%)	34 (6.4%)	17 (4.3%)	165 (8.8%)
PAU	103 (26.4%)	83 (15.1%)	89 (16.8%)	37 (9.4%)	312 (16.7%)

reading), and 4th-graders having more insertions and deletions (due to faster reading). Furthermore, for pseudowords, mispronunciations rise well above the defined false start type (PRE) since there are fewer attempts to correct unknown words, i.e., there is less awareness of errors done to try to correct them. Surprisingly, children did not use filled pauses when trying to read aloud as teen and adults do in spontaneous speech (Veiga et al., 2012), using silent pauses instead when halting their reading.

2.6.2 Reading speed

With annotated data, a simple analysis of the reading performance of each individual child can be done. A common metric is to evaluate reading speed considering only correctly read words, which is defined as Correct Words Per Minute (CWPM) (Hasbrouck and Tindal, 2006). The average values of CWPM per grade of annotated children are shown in Table 2.4, side-by-side with the target curricular goals (Buescu et al., 2015). A large inter-grade overlap of the distributions is observed, showing a variability in reading performance of different children, although the average does increase per grade. Figure 2.6 displays this behavior with a boxplot of the distributions of CWPM, showing one clear outlier for the third grade.

Table 2.4: Per grade mean and standard deviation of measured Correct Words per Minute (CWPM), Curricular Goals (CG) of CWPM and relative difference of CWPM to CG, for sentences and pseudowords reading tasks.

	Words in Sentences			**Pseudowords**		
Grade	***CWPM***	***CG***	***CWPM-CG***	***CWPM***	***CG***	***CWPM-CG***
1st	56.9±18.6	55	3.5%	18.2+-7.4	25	-27.4%
2nd	84.3±24.9	90	-6.3%	26.4+-8.4	35	-24.4%
3rd	95.1±24.7	110	-13.5%	25.6+-6.7	-	
4th	109.8±23.6	125	-12.2%	33.9+-9.1	-	

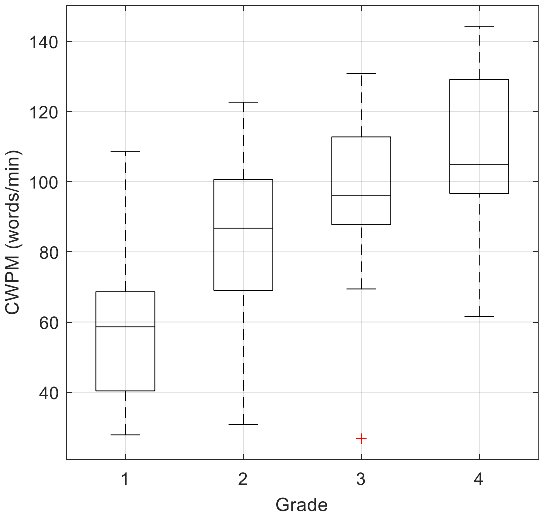


Figure 2.6: Median and quartiles boxplots of Correct Words per Minute (CWPM) for sentence reading tasks for each of the four grade levels. The cross sign indicates an outlier.

On data of adult speakers reading (Pellegrini et al., 2013), words per minute average 130.3±17.8. Comparing these values to the observed child performance, there may still be expected improvement from 4th grade children, although some perform as well as adults. For sentence reading, the difference from average CWPM to curricular goals increases in absolute terms along the grades, and these lower CWPM values may be explained by the difficulty of the reading tasks. It can be concluded that the suggested increase of difficulty along the grades may be too steep to directly evaluate CG as intended, and, for overall reading ability evaluation, this difficulty needs to be taken into account. For pseudowords, although there are no CGs for the third and fourth grades, average CWPM values are significantly lower than the CG, suggesting that the generated pseudowords (by joining common syllables and not based on existing words) are of high difficulty.

The defined curricular goals can be a starting point to appraise a child's reading ability. However, they do not take into account factors such as task difficulty or type of disfluencies; therefore, other ways to qualify reading performance should be considered.

2.6.3 Pseudoword performance and reaction time

To further analyze children's performance on the task of reading individual pseudowords, data from 100 children is considered, in which they read 10 individual pseudowords, one at a time per recording. This task differs substantially from sentence reading as morphological and phonemic awareness are the factors that influence a good performance on reading unknown words. Several interesting metrics can be extracted here, which may contribute to overall reading performance. First, the reaction time of starting to read the word (the time between the start of presentation and the onset of speech) reflects how fast the child achieves confidence in reading the entire word or the first syllable, especially for first graders. However, this metric does not reflect whether the word is read correctly or not, and there are children with fast reaction times who do make several mistakes. Still, the average reaction time decreases along the grades, as observed in Table 2.5 and Figure 2.7, with only a small increase from third to fourth grades (Proença et al., 2017b).

Also in Table 2.5, the number of words that had any disfluency event is indicated. For the first grade, the average of 6.5 disfluent words out of 10 is much higher than for other grades. Note that this measure is not identical to number of incorrect words (also presented in Table 2.5), since phone extensions or intra-word pauses may occur.

Table 2.5: Mean and standard deviation per grade of pseudoword reading reaction times (in seconds), number of uttered words with any kind of disfluency event (including extensions and intra-word pauses) and number of incorrect words.

	1st grade	**2nd grade**	**3rd grade**	**4th grade**
Reaction Time (s)	1.65±0.83	1.35±0.43	1.14±0.23	1.19±0.35
Number of disfluent words (out of 10)	6.54±2.89	3.23±2.32	2.96±1.87	2.70±2.24
Number of incorrect words (out of 10)	4.29±2.33	2.31±2.06	2.19±1.57	2.17±1.92

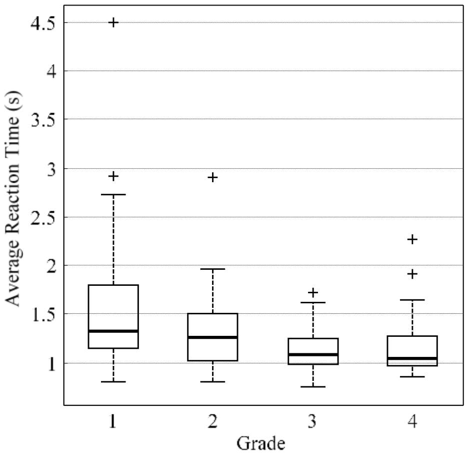


Figure 2.7: Median and quartiles boxplots of average Reaction Times for the pseudoword reading task for each of the four grades.

To analyze if the difficulty parameter used is suitable, the number of disfluencies given by children can be compared with the computed difficulty. Figure 2.8 shows a histogram-like distribution of the average number of disfluencies given per pseudoword through intervals of the difficulty parameter (Proença et al., 2016a).

An increase of disfluencies with higher difficulties is evident, showing that the parameter seems to be adequate. Due to the annotated data not presenting many repeated samples of the same tasks (resulting in, e.g., some pseudowords of high difficulty presenting zero disfluencies on three utterances), the correlation between difficulty and average number of disfluencies cannot be accurately computed.

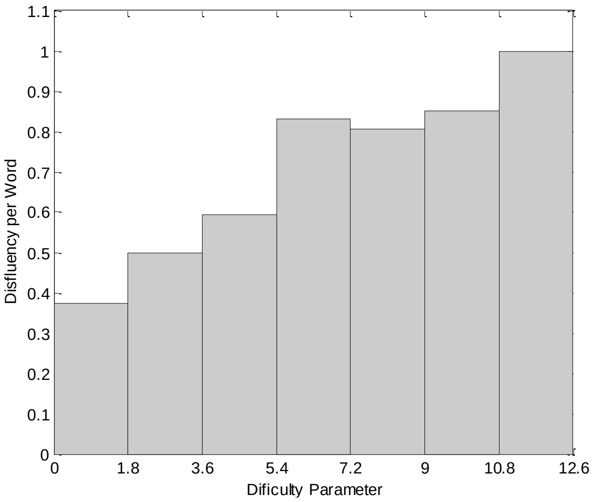


Figure 2.8: Average number of disfluencies per pseudoword for different difficulty parameter intervals.

Chapter 3

Automatic Detection of Disfluencies

Automatically detecting all the different types of disfluencies encountered in children's reading can be a significant challenge. In this work, the most frequent types are targeted for automatic detection – mispronunciations, false starts and repetitions – as well as intra-word pauses that can occur simultaneously with other events. It could be argued that detecting only correct words is enough to characterize children's reading in most applications, for example, to calculate correct words per minute. In that case, an approach such as word spotting (Veiga et al., 2014) could be enough to detect correct words, although it is expensive, especially for large sentences and would have to deal with repetitions. However, it has been found that features other than correct words per minute, such as the relative amount of specific disfluencies, may also be a relevant parameter for reading level assessment (Proença et al., 2017c) and the strategies developed stemmed from that goal. This chapter describes the effort of automatically detecting the targeted disfluencies, which will ultimately result in an automatic transcription of read utterances.

3.1 State-of-the-art

Although this study considers children aged 6 to 10 reading in their primary (native) language, the area of reading analysis is also relevant for second language learning. However, automatic analyses of foreign language reading (Abdou et al., 2006; Cincarek et al., 2009) are mostly targeted at adults or young adults for whom speech technologies are significantly more mature. Moreover, although similar reading problems may be encountered for young and adult readers, most problems in young children arise from the inability to follow phonological and phonotactic rules, as well as hesitations, self-corrections and slow reading speed. It is rarer to find problems of badly realized vowels, often the case in second language reading.

There are several known methods to detect disfluencies, such as those based on hidden Markov models (HMMs), maximum entropy models, conditional random fields (Liu et al., 2005) and classification and regression trees (Medeiros et al., 2013), though most of past research focusses on spontaneous speech. Applicability to read speech may not be straightforward since disfluencies vary according to different speaking styles (Moniz et al., 2014). Disfluencies in reading have different nuances, and some prior work has targeted the automatic detection of these events in children's reading, with the most relevant contributions described below.

Black et al. (2007) aim to automatically detect disfluencies in isolated word reading tasks. They found that human evaluators rated fluency as important as accuracy when judging reading ability. The target of detection is mostly sounding-outs, where a child first reads phoneme by phoneme (which can be whispered) and then reads the complete word. They trained HMMs and built a grammar structure specialized for disfluencies, capable of detecting partial words and allowing silence or noise between phones. The correct word is compulsorily considered to be pronounced in the final state of the grammar. The results achieved were 14.9% miss rate and 8.9% false alarm rate for the detection of hesitations, sound-outs, and whispering. By comparison, in the LetsRead data, no phoneme-by-phoneme sounding-out was found. Instead, there are syllable-by-syllable sounding-outs with possible silence between syllables, which will be addressed.

Duchateau et al. (2007) also target the reading of isolated words. Based on HMMs, they use a two-layer decoding module, first with phoneme decoding using phoneme-level finite state transducers to allow false starts with partial pronunciations, and then a second lattice to allow for repetitions or deletions of words. For the detection of reading errors on word reading, a miss rate of 44% and a false alarm rate of 13% were achieved. For a pseudoword reading task, they achieved a 26% rate of both misses and false alarms. In Yilmaz et al.

(2014), an extension to the work by Duchateau et al. (2007) is developed. The new evaluation is on a mixture of word and sentence reading tasks, and the models are still based on HMMs. The decoding scheme is more flexible to allow for the most common substitutions, deletions and insertions of phones in the language, as described by a phone confusion matrix. This confusion matrix was obtained by comparing the output of the recognizer with transcripts on a larger corpus. The final result for the detection of all disfluencies (word repetitions, stuttering, skipping and mispronunciations) was 44% miss rate at a 5% false alarm rate.

Hagen et al. (2007), targeting partial word pronunciations, found that syllables were the best subunits to use in a decoding lattice to detect these events. A 34.6% detection rate of partial words is achieved for a 0.5% false alarm rate, and the overall word error rate was similar to using a decoding grammar based solely on words.

Li et al. (Li et al., 2007) aim to track children's reading of short stories for a reading tutor. As a language model, they employ a word level context-free grammar for sentences to allow some freedom in decoding words. Each word also had a concurrent garbage model with the most common 1600 words, which allows detecting word level miscues, but was also able to detect some sub-word level miscues for short words. On a detection task of all reading miscues (including breaths and pauses), they achieved a miss rate 23.07% at a false alarm rate of 15.15%.

It should be mentioned that much of the prior research focuses on individual word reading tasks – exceptions being (Li et al., 2007) and parts of (Yilmaz et al., 2014) –, whereas the present work targets the reading of sentences and pseudowords. Unfortunately, it will not be straightforward to make comparisons of the results in literature to the obtained ones, since the data is significantly different and the targeted disfluencies also vary. Some studies go further and attempt to provide an overall reading ability index that should be well correlated with the opinion of expert evaluators (Black et al., 2011; Duchateau et al., 2007), which is also a direct application of the presented work, as discussed in the next chapter.

3.2 Two stages approach

In order to be able to extract several features to improve the classification of words as mispronounced or not, the developed strategies of disfluency detection are split into two stages:

1. First, a segmentation stage gets candidate word pronunciations along with their time-alignment information, while detecting word repetitions and false starts. A candidate word segment is ideally aligned with the pronunciation attempt of its word.
2. Secondly, a mispronunciation classification stage takes each candidate word segment, computes several features relating to the likelihood of pronunciation and phonetic sequence comparison and classifies the word attempt as correctly pronounced or not.

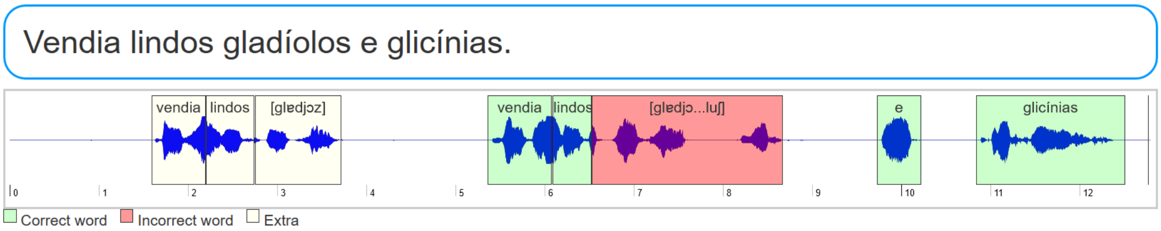
The workflow of disfluency detection with the final goal of detecting mispronunciations follows the schematic of Figure 3.1. Although the second stage is called mispronunciation classification, it is in essence a detection task, since candidate segments must be detected automatically (as attempted in the first stage). By using segmentation information from a manual transcription, the task can be considered a case of classification.

Utterance
Word Alignment / Candidate Segmentation
Mispronunciation Classification

Figure 3.1: Schematic of disfluency detection workflow.

Figure 3.2 shows an example of problematic reading with some of the disfluent events that need to be automatically detected. The figure also demonstrates the two stage process: getting segments and classifying the pronunciation.

Although incorrect intonations of a word may relate to incorrect reading, this work only focuses on deviations from the ideal phonetic pronunciation. The manual transcription of the LetsRead database includes phonetic deviations but no marking is done for intonation problems. In fact, the features used for the automatic methods are derived from speech



Vendia lindos gladíolos e glicínias.

Figure 3.2: Example of a prompt (top) and an utterance of its reading attempt (bottom), with segmentation for extra content and correct and incorrect words. Expected correct reading: [vẽdˈiɐ] [lˈĩduʃ] [glɐdˈiuluʃ] [i] [glisˈinjɐʃ]. Transcription: REP(vendia) REP(lindos) PRE([glɐdjɔz]) vendia lindos SUB([glɐdjˈɔ...luʃ]) e glicínias.

recognition/decoding paradigms. Correct prosody is linked to a good reading performance and it only be partially addressed by considering duration metrics. As previously stated, it has been shown that considering different types of disfluency rates in addition to reading speed features (without other prosody metrics) can already improve the prediction of a child's overall reading level (from 0.92 correlation for correct words per minute to 0.95 using multiple features) (Proença et al., 2017c, 2017b). The output of the developed methods is also a full automatic annotation of children's read utterances.

Several acoustic models were trained throughout the developed work but two are ultimately used for disfluency detection, using the training set of 9 hours:

- Triphone GMM-HMM models trained with the Kaldi toolkit (Povey et al., 2011). Used for the segmentation stage, the models have shown a satisfactory performance for decoding that is close to forced alignment.

- Phonetic recognizer based on long temporal context (FIT, 2015). The recognizer achieves 27% phone error rate on the test set with a free phone-loop model, which is an acceptable performance for the amount of training data. The 9 hours of data are insufficient to build satisfactory deep neural networks. These models are used to get information of non-speech frames as well as phone likelihoods and phonetic sequence recognition. This training was based on the work developed for the topic of Query-by-example spoken term detection (Proença et al., 2015e; Proença and Perdigão, 2016a, 2016b), as stated in the introduction.

3.3 Detection of Repetitions and False Starts

This first stage aims to get the best alignment possible for both correct words and mispronounced attempts, with the only given meta-information being the original prompt. The first challenge for segmentation comes from all the extra content that can frequently occur (repetitions, false starts and insertions) as well as deletions. Otherwise, a forced alignment of the prompt word sequence would suffice. The second challenge is that pronunciations can differ significantly from the reference pronunciation. Any decoding strategy must not be too unconstrained, since short words might otherwise be detected in false starts and mispronunciations. Consequently, the proposed decoding grammars are a mixture of strictly following the prompt with the added option of word repetition. It is still possible to obtain a good alignment for a mispronunciation, even if it deviates substantially from the reference word.

For acoustic models, greater success was found in this segmentation task by simply using triphone hidden Markov models of Gaussian mixture models (GMM-HMMs), rather than neural networks for decoding with triphones. One possible explanation is that the amount of training data (9 hours) is not sufficient for neural network training of robust models for word decoding.

Another substantial challenge is the occurrence of intra-word pauses. These are more common for the lower grades and stem from reading a word syllable by syllable, with a significant pause in between. A regular word decoder or aligner is not expecting silence inside of a word. Two ways to deal with intra-word pauses were investigated: cutting silence frames (more exactly non-speech frames) and decoding with word-based grammars, or using decoding grammar at the subunit level (syllables) and allowing silence between each subunit.

For this segmentation stage, the approaches taken initially will be described (which had significantly worse results than later approaches), but the ones ultimately described and compared in depth are a baseline and two methods that deal with intra-word pauses before or during decoding:

1. A baseline system where intra-word pauses are not considered and a word-based decoding is used, allowing repetitions and false-starts.
2. A word-based system where silent segments are cut in the waveform before decoding with full words, similarly to the baseline. This system was first reported in Proença et al. (2017a).
3. A syllable-based decoding grammar, where silence is optional between all syllables. This method was first reported in Proença et al. (2018).

The methods are described in the following subsections. For all methods, the allowed false starts (represented by the suffix PRE) are based on stopping a word pronunciation at syllable boundaries, which are common interruption points. For example, for a word with four syllables [syl1.syl2.syl3.syl4], the allowed pronunciations for a false start are [syl1], [syl1.syl2] and [syl1.syl2.syl3]. Allowing deletions was found to provide worse alignment results from the onset of this study. This is probably because introducing ‘skip’ arcs in the alignment lattices produces a high degree of freedom that allows mispronounced words to be matched up incorrectly, as previously stated. Therefore, deletions were not allowed in the presented approaches. Insertions of spoken content not related to the prompted words (including out-of-vocabulary words) were also not considered, but it is likely that candidate segments for insertions and deleted words (with an imperfect alignment) will be detected as mispronounced on the pronunciation classification stage. In fact, repetitions and false-starts represent 92% of extra events in the transcribed data.

3.3.1 Initial approaches

It might be relevant to report on the two initial methods developed to solve word segmentation, even if newer approaches improved results substantially. The overall strategy was the same throughout: cut silent segments and allow for word repetitions and syllable-based false starts.

For the first approach, reported in Proença et al. (2015d, 2015b), HMMs were trained with 3 hours of child speech using the HTK toolkit (Young et al., 2006). To deal with intra-word pauses, silence is cut at the Mel-frequency cepstral coefficient (MFCC) level, for segments of low energy of at least 220ms duration. For each sentence, a specific word decoding lattice was employed, allowing false-starts and repetitions. Figure 3.3a shows the sub-lattice schematic for each individual word and Figure 3.3b shows the sentence lattice. Realizations of speech are made on nodes and not on arcs. The probabilities/weights of PRE and repetitions were defined as 1%. Even if silent segments were cut, silence/non-speech (pause) is still allowed between each event. Repetitions of sequences of two or three words are also allowed at the sentence lattice level.

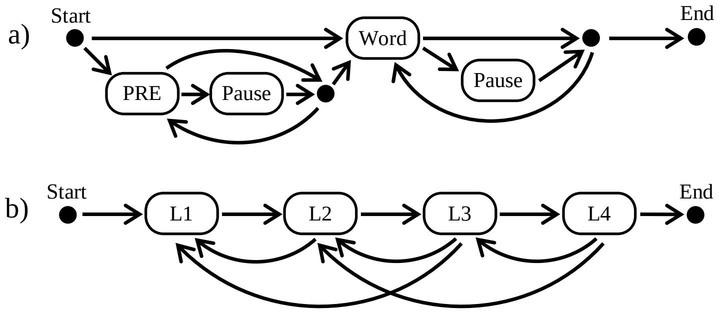


Figure 3.3: Schematic of an individual word sub-lattice (a) and example of a final lattice for a 4 word sentence using 4 word sub-lattices L1 to L4 (b).

Considering only words and extra events, the system results in 5.20% word error rate (WER) compared to 8.14% WER of a forced alignment of the original prompt. The alternative use of acoustic models trained with 40 hours of adult speech and adapted using 3 hours of child data with Maximum a Posterior (MAP) adaptation is also analyzed. The effect of varying word insertion log probabilities can be analyzed by detection error tradeoff (DET) curves (Martin et al., 1997). The DET graph usually provides clearer system comparison than receiving operator characteristic (ROC) curves by focusing on the area of interest and making a non-linear transformation which results in curves approaching a linear representation. Figure 3.4 shows the results of detection of events of false-starts and repetitions, comparing the use of the two acoustic models.

For approximately 5% false alarm rate, the system achieves 15% miss rate with child-only models and a 25% miss rate with adapted models. For a 1% false alarm rate, the miss rate falls around 30%. A conclusion taken early on from this study was that it was preferable

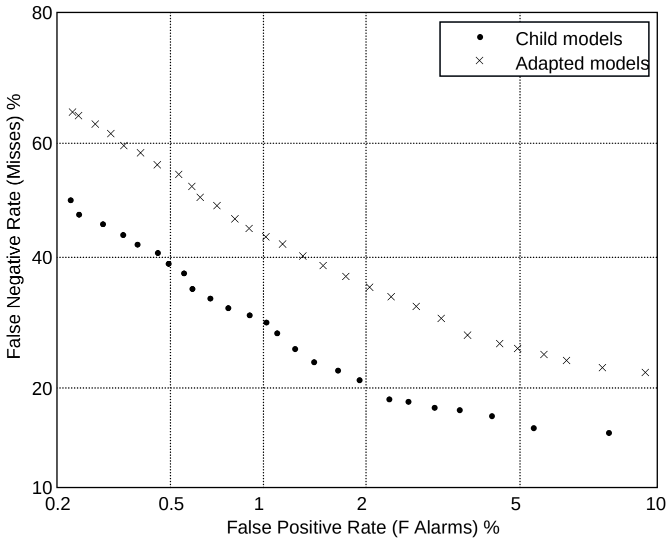


Figure 3.4: Detection Error Tradeoff (DET) curves for the first disfluency detection system with varying word insertion log probabilities, for two acoustic models.

to train acoustic models with child data only. Therefore, no further efforts were taken to build acoustic models with larger amounts of adult speech and adapt to children.

For the second approach, the Kaldi toolkit (Povey et al., 2011) was now used to train acoustic models and perform decoding. This method was reported in Proença et al. (2016b, 2017b). The training set of transcribed data was now 5h30m, and a test set of 1h30m was used. Low energy segments longer than 150ms are also cut to deal with intra-word pauses, but before parameterization, at the waveform level. The task-specific decoding lattices changed substantially: finite state transducers (FST) are used where realizations are defined by arcs and not nodes. At this stage, it proved challenging to build decoding-ready FSTs for all sentences with a high degree of freedom (it was later found that the problem could be solved for some sentences by skipping determinization). Therefore, any back-transitions that complicate the FST and increase decoding times are avoided, and this strategy remains valid as a solution for fast decoding. As an example, Figure 3.5 describes the FST grammar for the three-word sentence *ele sonhava muito* [ˈelə suɲˈavɐ mˈũj̃tu] (ˈhe dreamed a lotˈ). There is a basic group of FST nodes representing each word, allowing false starts and repetitions of the word. For example, the word ˈeleˈ is represented by the nodes 0, 1 and 2, and the transitions arriving at nodes 1 or 2. The same applies to the word ˈsonhavaˈ with nodes 2, 3 and 4, and for the word ˈmuitoˈ with nodes 6, 7 and 8.

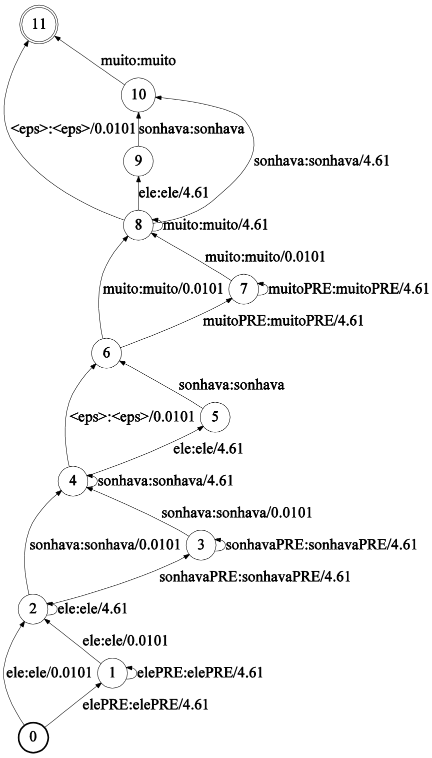


Figure 3.5: Sentence FST without back-transitions for the three-word sentence *ele sonhava muito* [ˈelə suɲˈavɐ mˈũȷ̃tu].

In addition, the possibility of repeating sequences of words is allowed, at most of the previous two or three words and only once for the same pattern. In the example FST, paths that go through nodes 5, 9 and 10, represent these possibilities. Furthermore, following the left-most arcs, one gets the original sentence without any false starts or repetitions (<eps> is an epsilon arc, consuming no input or output).

The best WER obtained on the training set was 3.75%, corresponding to a false alarm rate of 0.89% and a 23.53% miss rate. The DET curve of the training set is shown in Figure 3.6. Using the word insertion penalty and rescoring weight from the best WER, the results on the test set are: 4.47% WER, 1.94% false alarm rate and 20.60% miss rate. The best possible WER would be 4.01% by varying word insertion penalty.

This method was the one used for an automatic transcription of data as a starting point for the second stage of manual transcription, combined with an automatic mispronunciation classification of words based only in one feature (LLR-spotter) which will be described in

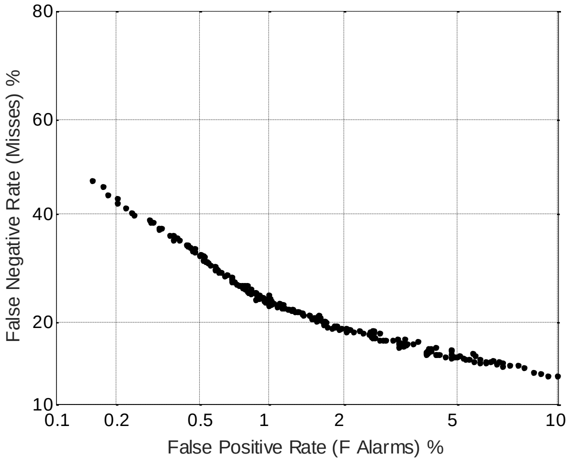


Figure 3.6: Detection Error Tradeoff (DET) curve using lattices without back-transitions for the detection of false starts and repetition events on the training set.

section 3.4. This was also the automatic transcription used for the study of automatic reading level assessment (chapter 4). Although improved results were being obtained by moving to Kaldi, the lattice was still very strict and did not represent all the possibilities observed in the data. Subsequent approaches, which will be more thoroughly analysed in the next subsections, will provide more freedom for repetition through simpler decoding lattices, or even solve the intra-word pause issue by employing subunit decoding.

3.3.2 Baseline

As a baseline approach, only repetitions and false starts are targeted and nothing is done to address intra-word pauses. For a given utterance, a decoding grammar (lattice) is built from the original prompt, allowing repetitions and syllable-based false starts. Decoding is performed using this lattice and HMM models.

The lattice built for a specific prompt is a finite state transducer (FST) based on the sequence of words of the original prompt. For each word, two additional elements are added to the lattice: an arc to go back after a word pronunciation, allowing for repetitions; and a self-loop arc before the word to allow multiple false starts. The arc weights of the FSTs used in this work were empirically decided. An example for the lattice built for the prompt *"ele sonhava muito"* [ˈelə suɲˈavɐ mˈũj̃tu] (he dreamed a lot) is shown in Figure 3.7. False starts are represented by the suffix PRE and, in this example: *elePRE* can only be [e]; *sonhavaPRE* can be [su] or [su.ɲˈa]; *muitoPRE* can only be [mũj̃].

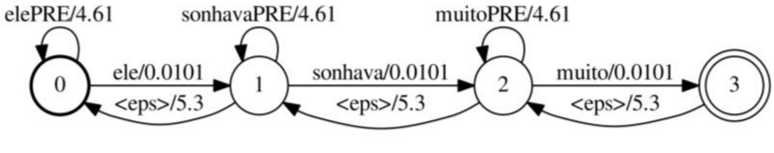


Figure 3.7: Schematic of the lattice built for the prompt ˈele sonhava muitoˈ, for the word based-decoding method.

Following the horizontal left-to-right arcs, the original sentence is obtained. By following multiple arcs that transition backwards (non-consuming <eps>), repetition of sequences of words are also allowed, such as *“ele sonhava ele sonhava muito”*. These occurrences are frequent in the data and typically represent corrections initiated by repeating every word from the start of sentence or clause.

3.3.3 Word-based decoding with silence cutting

The word-based approach reported in Proença et al. (2017a) is similar to the described baseline, but silence periods are removed before decoding so that words are expected to be continuous. The trained neural network based on long temporal context is used to provide posterior probabilities of phones and non-speech events. This output is used in this method to detect non-speech segments (it will also be used for mispronunciation classification). The method follows these steps:

1. Voice activity detection. Significant non-speech segments (longer than 150ms) are cut from an utterance to deal with intra-word pauses. Non-speech segments are found from sequences that have a high probability of being silence based on the posterior probabilities output by the phonetic recognizer.
2. Decoding using task-specific grammars. For a given utterance, the same word-based decoding as described for the baseline is employed, allowing repetitions and false-starts (Figure 3.7).
3. Reintroduction of non-speech segments. Finally, information pertaining to the segments of non-speech that were originally cut (either separating words or inside

a word) is used to expand the decoded segmentation to the utterance’s original duration.

3.3.4 Syllable-based decoding

In the final approach the problem of silence inside a word is handled differently. Here, the decoding strategy is based on separating a word into its syllable components and building a lattice solely with these syllables. Figure 3.8 shows an example of the lattice for the same sentence “ele sonhava muito” [ˈelə suɲˈavɐ mˈũj̃tu]. The allowable repetitions and false starts are similar to the previous approach since, after a given syllable, it is only allowed to return, at most, to the beginning of the word.

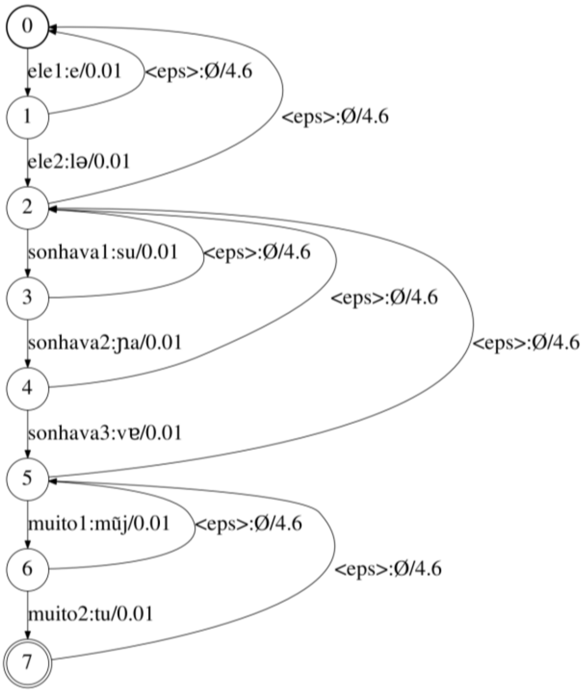


Figure 3.8: Schematic of the lattice built for the prompt “ele sonhava muito”, for the syllable based-decoding method. Optional silence is implicitly allowed at every node. Considering ele2:lə/0.01 - ele2 is the second syllable of the word ˈeleˈ, with pronunciation [lə] and negative log-probability of 0.01).

Although it is not depicted, optional silence is allowed at every node. Sequences of words can also be repeated, as there are continuous back-transitions for full words. Although not shown in the example, if multiple pronunciations are possible for a given word, they are taken into account as alternate pronunciations of a syllable. After decoding, a reconstruction step is needed to join adjacent syllables into their corresponding word, repetition or false start.

The output of these approaches is a per-utterance segmentation into word-relevant segments, be they false starts, repetitions or word-candidates, to be classified in the next stage as mispronounced or not.

3.3.5 Results

Using both the word-based decoding with silence cutting and the syllable-based decoding for the segmentation stage, the overall word error rate (WER) and the detection of extra speech events (repetitions and false starts) can be analyzed.

WER is analyzed by comparing the decoded sequence of words and events to the reference transcription. Using the full text of the original prompts as hypothesis, WER is 9%, where errors correspond to repetitions, false starts, insertions and deletions. Since the decoding strategies do not take insertions and deletions into account, these will always appear as errors. WER values for the test set, when using the optimal insertion penalty for the training set, for the baseline without cutting silence and both segmentation approaches are presented in Table 3.1, as well as results when using the optimal penalty for the test set. The values on Table 3.1 for the best results on the test set (best) are just to get the hypothetical minimum/best results, by using the best possible insertion penalties for the test set. This shows how close the results are to applying the training set operating points (insertion penalties) on the test set vs. a theoretical optimization on the test set, which indicates if the operating point is correct. These optimizations on the test set are not used further on.

Table 3.1: Overall word error rate (WER) and miss and false alarm rates for the detection of repetitions and false-starts, for the test set, using the optimal insertion penalty of the train set and the best one on the test set (best).

Segmentation approach	**WER %**	**Miss %**	**False Alarm %**	**WER % (best)**	**Miss % (best)**	**False Alarm % (best)**
Baseline	4.15	15.92	2.36	4.11	17.60	2.21
Word-based w/ sil. cut	2.45	11.17	0.98	2.41	10.61	0.98
Syllable-based	2.17	11.45	0.68	2.16	10.89	0.70

The recognition of repetitions and false starts is considered as a detection task, lumping both into a single class. Although the false starts allowed are up to the last syllable, in the transcribed data some are complete mispronunciations of a word, with a subsequent attempt of correction. Those are possibly detected as repetitions with these methods and motivate analyzing the detection of both repetition and false start events as just one class. To evaluate a system's performance in the detection of these events, it is stipulated that:

- Extra events (insertions) are false alarms;
- Undetected events (deletions) are misses;
- Events detected as belonging to a different word (substitution) are also misses. For example, a false start of one word may be detected as a repetition of the previous one.

These specifications are similar to the ones used in NIST evaluations (Fiscus et al., 2007). However, to calculate false alarm rates, the number of false alarms are divided by the total number of original words. Figure 3.9 shows the detection error trade-off (DET) for the segmentation approaches on the test set, obtained by using a wide search beam during

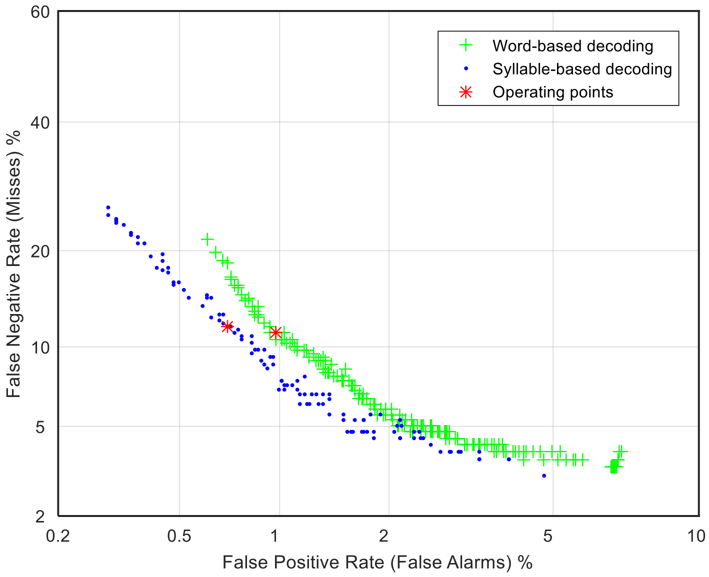


Figure 3.9: Detection error tradeoff (DET) for the detection of repetitions and false starts on the test set, for both decoding approaches. Operating points are with the best train insertion penalties.

decoding and various word insertion penalties and lattice rescoring weights. Operating points correspond to using insertion penalties and lattice weights that resulted in the best WER on the training set. The equal error rate (EER) for the systems is not of interest since it corresponds to relatively high false-alarm rate (equal to the miss error rate) and higher WER, far from the targeted operating points. Table 3.1 presents the resulting miss and false alarm rates at these points, as well as the best possible ones by optimization on the test set.

Comparing the WER results as well as the DET values, it is clear that the second approach – syllable-based decoding without cutting silent segments – performs better than the alternative method of cutting silent segments and aligning full words. Still, the above results do not take into account the time-wise alignment information. Comparing the decoded hypotheses to the manual transcription (reference), three metrics for alignment match are analyzed and presented in Table 3.2. The metrics are:

- overlapRef – percentage of frames that overlap, per event, over the length of the reference word, averaged for all events.
- overlapOverMax – percentage of frames that overlap, per event, over the maximum interval between the reference and hypothesis start and end frames, averaged over all events. This metric penalizes longer hypotheses.
- overlapUtt – percentage of overlap frames per utterance over the length of all events of the sentence, averaged over all utterances.

For each event of the reference transcription, only one hypothesis is compared: the one that corresponds to the same word and with the largest overlap if more than one exists. Therefore, all the metrics penalize shorter hypotheses.

Table 3.2: Metrics analyzing frame-wise alignment match between the manual transcription and decoded hypotheses (defined in the text).

Segmentation approach	**overlapRef %**	**overlapOverMax %**	**overlapUtt %**
Baseline	89.00	82.52	90.26
Word-based w/ sil. cut	89.71	84.42	91.98
Syllable-based	90.21	83.71	92.68

The alignment metrics also show that the syllable-based decoding performed better overall, although a smaller overlapOverMax shows that it may have provided slightly larger segments. Subsequently, only this automatic segmentation method will be used to analyze mispronunciation classification results. The method has the advantage of skipping a non-speech removal step that needs to decide a minimum duration for non-speech segments to be cut. Leaving this decision as optional silence in the decoder seems to be ideal.

3.4 Detection of Mispronunciations

The analyzed reading material is extensive, including challenging words and pseudowords. Therefore, the proposed approach targets the possibilities of mispronunciation in a general way, in contrast to considering typical pronunciation errors during decoding. Mispronunciations by children can range from a simple change in one phoneme, or changes in phoneme order, or phoneme deletion or insertion, to severe changes from the intended correct reading. Intonation problems are currently not targeted by this work.

A straightforward possibility to decide if a word pronunciation is correct or not is to compare the uttered sequence of phonemes to the allowable pronunciations. If there is a match, the pronunciation would be considered correct. However, the accuracy of automatic phoneme recognition is not high enough to support this method, since recognition inaccuracies (insertions, deletions and substitutions) can lead to numerous false alarms of mispronunciation. Therefore, methods based on the word likelihood given the correct pronunciation prove to be more successful. Phoneme recognition will still be applied to obtain additional inputs for mispronunciation classification. In fact, word pronunciations will be classified based on multiple individual features and combining them in multi-feature classifiers.

For this task, a neural network based on long temporal context (FIT, 2015) was trained, as mentioned in section 3.2. It is targeted for phoneme recognition and its output, used here as the basis of likelihood computations, are state-level posterior probabilities of 34 phones with 3 states each, including non-speech. Compared to the phonetic alphabet of Annex II, semi-vowels [j], [w], [j̃] and [w̃] are joined to its vowel counterparts [i], [u], [ĩ] and [ũ], since phonetic recognition results were slightly better compared to separating them. Also, no distinction is made for stressed or non-stressed vowels. An example of the obtained posterior probabilities per frame for the reading attempt of the prompt “o carro do meu tio é azul” is shown in Figure 3.10.

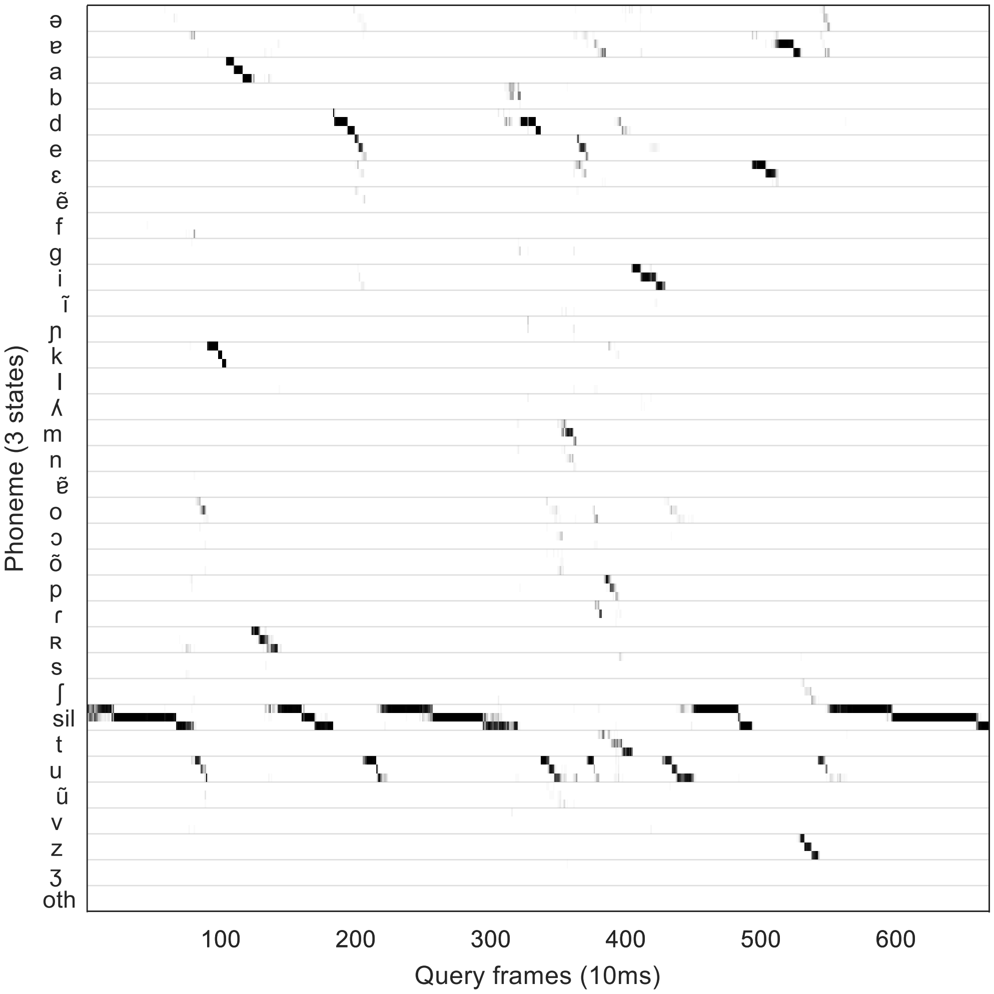


Figure 3.10: Phonetic posterior probabilities per frame for the utterance “o carro do meu tio é azul”. The child said “o carro deu do meu tio é azul”.

There are two considerable factors that need to be addressed to verify correct pronunciation: multiple allowable pronunciations due to commonly used pronunciation variants or accent-based variations; changes in pronunciation due to context and articulation with adjacent words (coarticulation). For all the features of mispronunciation classification that need to consider the reference pronunciation of a word, these possible variations need to be included. The multi-pronunciation rules defined are:

- For the phone sequence [ɐj], duplicate with [ej], or vice-versa, e.g. in the word <seis>: [sɐjʃ] or [sejS]
- For words beginning with graphemes <ex> followed by a consonant, allow both [əʃ] and [ɐjʃ], e.g., for the word <excerto>: [əʃsˈeɾtu] or [ɐjʃsˈeɾtu].
- For words ending with [o], allow [ow] as well, e.g., for the word <fechou>: [fəʃˈo] or [fəʃˈow].
- Allow the removal of instances of schwa [ə], since it is often unspoken or an artifact of vocalization of previous phones in regular speech, e.g., for the word <semente>: [səmˈẽtə], [smˈẽtə], [səmˈẽt] or [smˈẽt]. This is usually not a problem for word recognizers where triphones solve the issue of non-existing schwa, but may be problematic for the phonetic recognizers.
- Irregular cases of common word contractions that are usually not considered incorrect pronunciations: <para> with [pˈɐɾɐ] or [pɾɐ], <pelo> with [pˈelo] or [plu].
- For pseudowords, there can be words that may have several allowable pronunciations (like the case of heteronyms) and multiple pronunciations were defined manually, e.g.: <frecal> with [fɾɛkˈal] or [fɾəkˈal], <jodas> with [ʒˈodɐʃ] or [ʒˈɔdɐʃ].

Additionally and exclusively for words with pronunciations larger than 3 phones:

- For words beginning with [ə], allow [i] as well, e.g.: <escuta> can be [əʃkˈutɐ] or [iʃkˈutɐ] (or [ʃkˈutɐ] from the schwa rule above).
- For words beginning with [e] followed by a consonant, allow [i] as well, e.g.: <energia> can be [enəɾʒˈiɐ] or [inəɾʒˈiɐ].
- For words beginning with [o] or [ɔ] followed by a consonant, allow both [o] and [ɔ], e.g.: <origem> can be [oɾˈiʒɐ̃j̃] or [ɔɾˈiʒɐ̃j̃].

Context and word co-articulation rules are really only relevant for sentence reading, where allowable pronunciations can appear due to interactions with adjacent words. Following certain context rules, the issue was treated similarly to multiple pronunciation, by adding new possible pronunciations to be checked, case by case. The applied rules are based on the last phones of a word and the first phones of the following word, based on the phonetic dictionary. However, the time-alignment of the analyzed utterance is taken into account (e.g., automatic segmentation for section 3.3) and the first main rule is that nothing is done if the adjacent event is silence or non-speech. The rules defined for coarticulation are:

- For words ending in [ɐ] with the next word beginning with [ɐ], both words' phones can change to [a], e.g.: for <para aqui> [pˈɐɾɐ ɐkˈi], <para> can also be [pˈɐɾ**a**] and <aqui> can also be [**a**kˈi]. Although if the contraction is done the speaker only says one phone [a], it is added on both words since they are to be analyzed for correctness of pronunciation separately (and there will be a small leeway on the segmentation boundaries).
- For words ending in [ʃ] followed by a word beginning with a voiced consonant, [ʃ] can be changed to [ʒ], e.g.: for <os gatos> [uʃ ɡˈatuʃ], <os> can also be [u**ʒ**].
- For words ending in [ʃ] followed by a word beginning with a vowel, [ʃ] can be changed to [ʒ] or [z], e.g.: for <as aulas> [ɐʃ ˈawlɐʃ], <as> can also be [ɐ**z**] or [ɐ**ʒ**].

3.4.1 Individual Features

The proposed features for mispronunciation are introduced in this section and they will be taken into account for multi-feature classification models. Features are extracted for each candidate segment from the previous segmentation stage, representing a word. The features are based on both expecting that the correct pronunciation is there as well as on what can be recognized in that segment.

Goodness of pronunciation (GOP) (Witt and Young, 2000; Pellegrini et al., 2014) is a common metric to detect phonetic mispronunciations by computing the likelihood of a phone realization to belong to the ideal phone that should have been pronounced. For this stage, GOP-like features on phone posterior probabilities are computed, as well as edit distances of recognized versus ideal phone sequences and other details about the word. The best feature considers a log-likehood ratio (LLR), which is based on comparing the likelihoods of a keyword model and a filler model. More detail about LLR is presented on the LLR-spotter feature description below.

The computed features are listed in Table 3.3, with brief descriptions that are expanded in the text. Figure 3.11 shows an example of an aligned segment for a word, forced

Table 3.3: Main features considered for mispronunciation classification.

Feature abbreviation	Summary
LLR-spotter	Log-likelihood ratio based on word-spotting (also LLR-s)
LLR-ali	Log-likelihood ratio strictly over the original alignment
min-GOP	Minimum (worst) likelihood of a phone from a forced alignment
mean-GOP	Average likelihood of phones from a forced alignment
maxBadPhnProb	Maximum post probability of mismatched (bad) recognized phones
sumBadPhnProb	Accumulated post probability of mismatched recognized phones
LevBigram1	Levenshtein distance using a bigram phonetic model
LevBigram2	Same as LevBigram1 but with lower substitution weights for phonemes of similar phonetic class
LevBigram2	Same as LevBigram1 but with substitution weights according to a phone confusion matrix from phone recognition
LevPL1	Levenshtein distance using a constrained phonetic lattice
LevPL2	Same as LevBigram2 but using the constrained phonetic lattice
LevPL3	Same as LevBigram3 but using the constrained phonetic lattice
LevSumX	Three features for the sum of LevBigram and LevPL, where X: 1, 2, 3
LevProdX	Three features for the product of LevBigram and LevPL, where X: 1, 2, 3
Nframes	Number of frames from LLR-spotter
Area	Area of the LLR-spotter graph
Diff1	Difficulty of the word based on phonetic complexity rules (section 2.1)
Diff2	Same as Diff1 but without accounting for word length
OLD20	Mean edit distance of the word to its closest 20 orthographic neighbours
Nphones	Number of phones of the ideal pronunciation
Ngraph	Number of graphemes of the ideal pronunciation
Nsylls	Number of syllables of the ideal pronunciation
LLR-s/X	Normalizations of LLR-spotter where X: Nframes, Nphones, etc.
LLR-s*X	Interactions of LLR-spotter where X: LevBigram1, Diff1, etc.

alignment of phonemes, recognized phonemes from a bigram model, and how the features of LLR-spotter and LLR-ali are obtained.

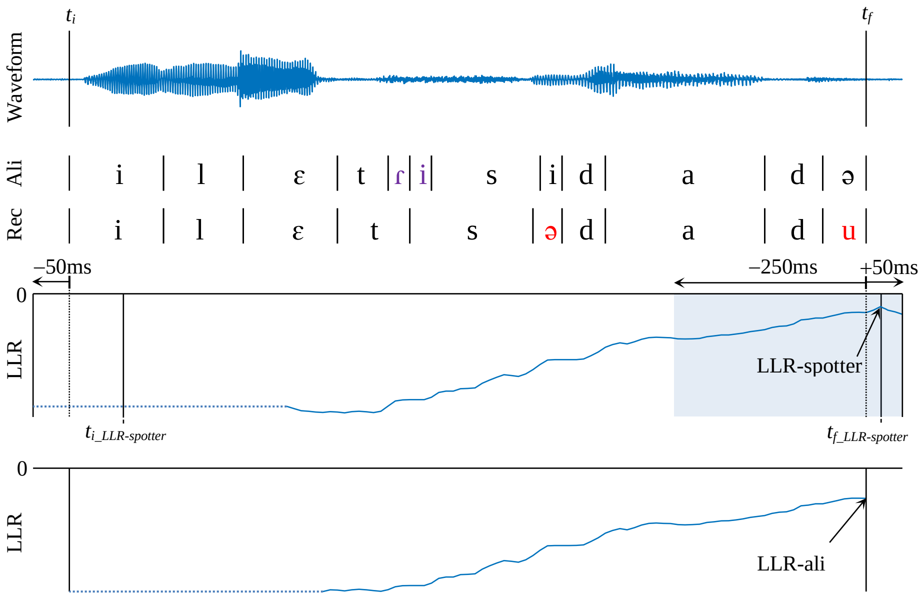


Figure 3.11: Schematic of an automatic alignment for the word *eletricidade* [ilɛtɾisidadə], with (top to bottom): waveform signal; forced alignment of reference pronunciation (Ali), recognized phones from a bigram model (Rec) resulting in a Levenshtein distance of 4 (2 deletions and 2 substitutions), LLR from a spotting approach (flexible beginning and end) and LLR with fixed beginning and end.

The features considered for mispronunciation classification are described below.

LLR-spotter: A GOP-like accumulated log likelihood ratio (LLR) from a word spotting approach (LLR-spotter or LLR-s for short). Although the previous stage provides alignments for candidate segments, these may not have the ideal boundaries to calculate likelihood, due to segmentation errors. This can even be the case if segmentation from manual transcription is used (e.g., including some silence inside marked boundaries). LLR-spotter is extracted from a word-spotting approach, as presented in Figure 3.12, in the near vicinity of the alignment boundaries (-50ms and +50ms). The keyword model is the sequence of ideal phones and the filler model is the free phone loop. Peak LLR between the models of ideal word and free phone loop is found in the vicinity of the ending time of alignment (-250ms and +50ms, empirically tuned), as shown in Figure 3.11. The best starting time is in the output token of the keyword model at each frame, as a result of the token-passing decoding approach (Young et al., 1997). The keyword model may also win at different starting times, and new boundary information is obtained. In essence, if only this feature were used, the purpose of the initial segmentation stage would be only to obtain approximate boundaries for candidates.

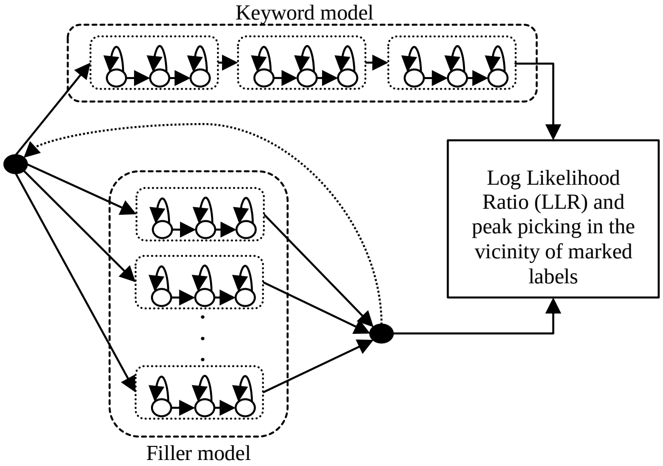

Keyword model
Log Likelihood Ratio (LLR) and peak picking in the vicinity of marked labels
Filler model

Figure 3.12: Schematic of obtaining a log-likelihood ratio from a word-spotting approach, using the keyword model of ideal pronunciation versus a filler free phone-loop model (Veiga et al., 2014).

LLR-ali: A log likelihood ratio based strictly on the original segment boundaries. LLR-ali is obtained similarly to the word spotting method, but the starting time has to be the initial frame and the ending at the final frame of the original segmentation as shown in Figure 3.11. Although it is considered mainly to compare to LLR-spotter, it might have alternative information for multi-feature classifiers. For example, LLR-spotter may find a words' correct pronunciation when in reality something extra was said (such as a plural), which may be balanced by LLR-ali if that extra was aligned for that word in the segmentation stage.

min-GOP and mean-GOP: Minimum and average GOP of phones, measured using a posteriori probabilities of phones from the phonetic recognizer neural network. For a forced alignment of ideal phones over the new interval from the word spotting method, the minimum (worst) likelihood of the aligned phones is obtained as a feature, as well as the average likelihood over all phones. It is expected that low likelihoods for reference pronunciation phones will indicate mispronunciation.

maxBadPhnProb and sumBadPhnProb: Maximum probability and accumulated probability of mismatched phones. As an approximately inverse idea to min-GOP, a free phone loop is used over the posterior probabilities to recognize the uttered sequence of phones. For each recognized phone that does not match the ideal pronunciation, the average posterior probability is taken over its alignment. Both the maximum and sum of these values are taken as features. It is hoped that a mismatched phone with very high probability from the phonetic recognizer will indicate an increased confidence that the word is mispronounced.

Levenshtein distances from phonetic recognition: Levenshtein distances (Levenshtein, 1966), or edit distances, are computed between the ideal phone sequence and the output of two phonetic recognition approaches, for each candidate segment:

- Bigram model. With improved recognition results over a free phone-loop model, a phonetic bigram language model is obtained from the training set and used to decode the best recognized sequence of phones over the candidate segment.
- Phonetic lattice (PL) based on ideal pronunciation. To overcome some errors by the recognizer’s output, constrained decoding models are built for each word, based on the notion that the ideal sequence of phones is the most probable to be detected on the segment. Loosely based on an implementation of a bigram model, a less probable back-off with a free phone loop is allowed in addition. The ideal sequence of phones has a higher probability, and only where deviations to this sequence are highly likely does the decoder choose the path of additional phones. An example of the phonetic lattice built for the word *azul* [ɐzul] (blue) is shown in Figure 3.13. The path through nodes 0-1-2-3-4-5-7 represents the correct word and those transitions are highly probable (low value of negative log probability). PHN represents all possible phones concurrently (self-loop at node 6), including silence. At the start or after a certain phone of the word, the most probable is the next correct one, as can be expected from a simple bigram model. Although posterior probabilities could also be obtained from the decoded lattices, they are often close to 1 due to the constrained decoding.

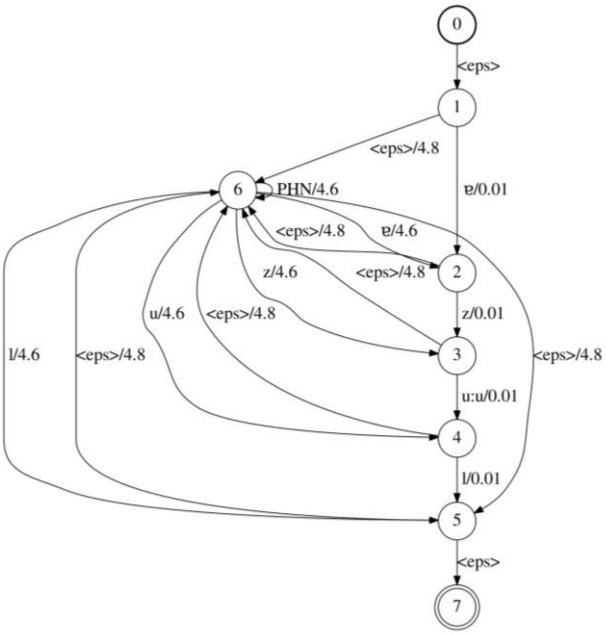


Figure 3.13: Example of the constrained phonetic lattice (PL) built for the word azul [ɐzul] (blue).

To compute the Levenshtein distance, phonetic substitutions, deletions and insertions are considered with the same unitary weight (cost). For example, ideal [ɐzul] versus recognized [ɐsul] results in a Levenshtein distance of 1 for the substitution of [z] for [s], whereas [zuiʃ] results in a distance of 3 for the deletion of [ɐ], substitution of [l] by [i] and insertion of [ʃ]. For each recognition, this distance is taken as a feature (LevBigram1 and LevPL1). Additionally, two slightly different distances are calculated: an edit distance with lower weights for substitutions among similar phonetic groups (LevBigram2 and LevPL2) where, for example, a substitution of [f] for [v] has a lower weight; an edit distance where the substitution weights are based on the phonetic confusion matrix from the output of the phonetic recognizer on the training set (LevBigram3 and LevPL3), following the suggestion of Yilmaz et al. (2014). For example, if the recognizer often detects [ɔ] for reference [o], this substitution will have lower cost.

Since it is expected that these two phonetic decoding approaches will have differing outputs, additional features are defined by combining the two edit distances, either by the sum or product of the values, for the three types of distances (LevSum1, LevSum2, LevSum3, LevProd1, LevProd2 and LevProd3).

Metrics from LLR-spotter: Based on the best spotted segment obtained in LLR-spotter, the detected number of frames (Nframes) and the LLR area (Area) are also included as features. Area is mostly used for normalizing LLR and is computed by summing the difference of LLR to the best LLR, frame by frame from the beginning to end of the best spotted segment (Veiga et al., 2014).

Difficulty and OLD20: Metrics of word difficulty. It is expected that harder or unusual words are more likely to be mispronounced. The difficulty of the word based on dubious and harder pronunciation rules as described in section 2.1 is considered with and without accounting for word length (Diff1 and Diff2). Additionally, the OLD20 metric of the word is considered, which is the mean Levenshtein distance of the word to its 20 closest orthographic neighbors from a large lexicon (Yarkoni et al., 2008), which may indicate a degree of unfamiliarity. As examples: the word ˈpratoˈ has an OLD20 of 1.45 due to 11 neighbors of distance 1 and 9 neighbors of distance 2; the word ˈpintalegretaˈ has an OLD20 of 5.75 due to 7 neighbors of distance 5, 11 neighbors of distance 6 and 2 neighbors of distance 7.

Word length: Additional features include the number of phones of the closest allowable pronunciation (Nphones), its number of graphemes (Ngraph) and number of syllables (Nsylls).

LLR normalizations and interactions: Several normalizations and interactions of LLR-spotter with other features are considered, by division or product, represented as, e.g., LLR-s/Nframes or LLR-s*LevBigram1.

3.4.2 Multi-feature Models

The goal of this stage is to classify whether a word is mispronounced or not and mispronunciations were defined as the positive class and correct pronunciations as the negative class. Therefore, the task is considered as a problem of binary classification. If only one feature is analyzed, its values will correspond to the discriminant for which detection error trade-off (DET) curves can be constructed, by using each discriminant value as a threshold for decision. For models that combine several features, it is preferred that new continuous values are outputted, as the new discriminants, and not a binary decision.

Although continuous outputs could be interpreted as degrees of correctness of pronunciations, this interpretation is not explored here.

To combine the information of several features, aiming to improve the classification of mispronunciations, the following models taking multiple inputs are investigated:

- Logistic regression (Logit). A logistic regression model for a binomial distribution, a case of generalized linear regression, was trained through maximum likelihood estimation. The logistic function (3.1) gives response probabilities by the linear combination of predictive features.

  $$\hat{y} = \frac{1}{1+e^{\mathbf{a}^T\mathbf{x}+b}} \tag{3.1}$$

  In (3.1), $\hat{y}$ is the predicted output, ranging between 0 and 1, corresponding to the probability of the sample being in the mispronounced class based on a linear combination of features where $\mathbf{x}$ is the feature vector, $\mathbf{a}$ is the coefficient vector (weights) of the input features and b is the intercept (bias) term.

- Neural networks (NNs). Networks were built with one hidden layer with variable number of neurons and one output, trained with Levenberg-Marquadt backpropagation (Hagan and Menhaj, 1994) and optimizing cross-entropy. The transfer function for the hidden neurons is the hyperbolic tangent sigmoid and at the output layer a logistic sigmoid function is used, providing output between 0 and 1.

- Support vector machines (SVMs). SVMs for binary classification were trained with a second order polynomial kernel, *C* parameter of 0.1 and an automatic heuristic kernel scale. To obtain continuous values, the considered output is not the binary decision but the classification score, which is the distance of the input vector to the decision boundary.

The hyperparameters for NNs and SVMs were chosen empirically. To avoid over-fitting to the training set, an alternative to using the entire set of features is analyzed. Stepwise feature selection is applied (Draper and Smith, 1998), through two approaches: adding features step by step when no features are included (Step-add) and removing features step by step when all features are included (Step-remove). For Step-add, the feature that minimizes deviance[17] when a logistic regression is applied was selected. However, a feature is only added if the decrease in deviance is statistically significant according to a chi-square test ($p < 0.05$). Similarly, for Step-remove, features are removed if, with their presence, the increase in deviance has a $p > 0.10$. This usually leads to different features being selected by the two approaches, with a logistic regression applied at the end. NN and SVM are again analyzed by using only the selected features as input (NN-step and SVM-step).

3.4.3 Results

Given candidate word segments with information about start and end times and corresponding prompt word label, an automatic classifier decides whether the word was mispronounced (positive class, 1) or not (negative class, 0). Using continuous values for predictions or probabilities of belonging to a class each output value can be used as a threshold for decision and derive DET curves.

Two sources of candidate segments will be analyzed: manual segmentation from the manual transcription and automatic segmentation from the syllable-based automatic decoding. It is expected that manual segmentations provide the best results and the clearest analysis of which features are important to classify mispronunciations. With automatic segmentation, some misalignments with the ground truth (manual segmentation) are to be expected. However, there must be an overlap of alignment in order for them to be considered a match.

Two groupings for mispronounced classes will be analyzed:

- SUB: only severe mispronunciations (SUB) as the positive class, without considering slight mispronunciations (PHO), since the latter are usually harder to detect.
- SUB+PHO: all mispronunciations as the positive class;

[17] Deviance can be computed by the sum of unit deviances given for a binomial distribution by $2\left\{y\log(y/\mu)+(1-y)\log\left((1-y)/(1-\mu)\right)\right\}$ with observation y and prediction μ (Song, 2007).

To compare classifications, given continuous output values for candidates, false alarm and miss rates will be analyzed. Since positive samples (mispronunciations) occur much less frequently than negative ones (correct words), the maximum accuracy measure would often relate to very low false alarm rates but with miss rates higher than 50%. Other measures such as F-score do not take into account the number of true positives. To target more interesting operating points, it is proposed to combine false alarm rate (FA) and miss rate in a weighted cost metric (3.2), where minimal cost is better.

$$Cost = w_1 \cdot FA + w_2 \cdot Miss \tag{3.2}$$

In (3.2), w_1 is the weight given to false alarms and w_2 is the weight of misses. If the weights were identical, the minimal cost would usually fall on an equal error rate. However, if more weight is given to false alarms, the optimal point will move towards fewer false alarms. It was decided to target a point that would not reach miss rates higher than 50% on the best results, but usually falls around 5% false alarm rate, achieved by defining w_1 as 1 and w_2 as 0.33.

The point of minimum cost in the training set will define the decision threshold to apply to test data. For all the considered groupings (manual and automatic segments, SUB and SUB+PHO as classes), two separate analyses of classifier model training and testing will be done:

- Cross-validation over the training set (CV-train). A 5-fold cross-validation is done using the training data used for acoustic models. The results are obtained by aggregating the outputs on the test data in each fold.
- Test. Predictions on the test set are made by training a model over the entire training set.

For models that depend on random initialization (NN weights and SVM automatic heuristic kernel scale), results will be of the model of minimum cost over 10 runs on training data.

Table 3.4 summarizes the results of the obtained cost metric when using only the individual features for classification of the SUB+PHO class. Features that are not shown, such as difficulty and number of phones, provide comparatively very poor results individually. Further normalizations and interactions provide either similar or slightly worse results compared to the displayed ones. The full table of results of individual features can be found in Annex III.

Table 3.4: Cost results for the classification of SUB+PHO class vs. correct words, using individual features.

	CV-train		**Test**	
Feature	**Manual**	**Auto**	**Manual**	**Auto**
LLR-spotter	**0.136**	**0.141**	**0.157**	**0.163**
LLR-ali	0.171	0.191	0.187	0.192
min-GOP	**0.194**	**0.200**	**0.234**	**0.233**
mean-GOP	0.246	0.251	0.276	0.273
maxBadPhnProb	0.280	0.248	0.288	0.276
sumBadPhnProb	**0.232**	**0.231**	**0.241**	**0.247**
LevBigram1	0.248	0.247	0.263	0.262
LevPL1	0.227	0.235	0.242	0.252
LevSum1	**0.199**	**0.206**	**0.212**	**0.221**
LLR-s/Nphones	0.156	0.168	0.187	0.190
LLR-s*LevBigram1	0.159	0.162	0.179	0.185
LLR-s*LevPL1	0.179	0.194	0.203	0.208

LLR-spotter proves to be the best performing feature, with a significant improvement over the similar LLR-ali metric, where the initial alignment is used, even for the manual segmentation. The effect of selecting multiple thresholds of feature values to classify candidates as mispronounced or can be analyzed by plotting DET curves. The comparison of the DET curves of LLR-spotter and LLR-ali is shown in Figure 3.14, where the improvement of LLR-spotter is clear. Figure 3.15 shows the DET curves of the next best 3 features (barring normalizations of LLR-spotter) – min-GOP, sumBadPhnProb and LevSum1 – showing different performances on varying false alarm rates. However, it is interesting to note that they achieve similar performances, given their separate origins.

LevPL1, the Levenshtein distance by using the constrained phonetic lattices (PL), provides a better cost metric than the bigram one (LevBigram1), with their combination proving successful (LevSum1). Analyzing the phone error rate (PER) for the two phone

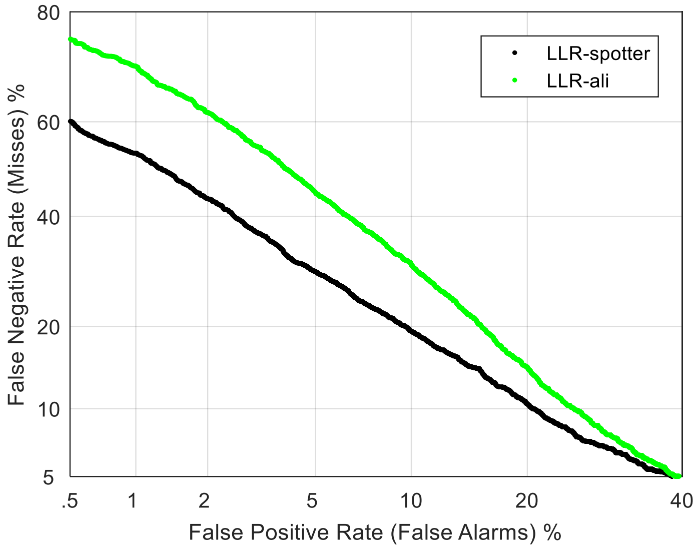


Figure 3.14: DET of LLR-spotter and LLR-ali as solo classification features, for SUB+PHO on CV-train, using automatic segmentation.

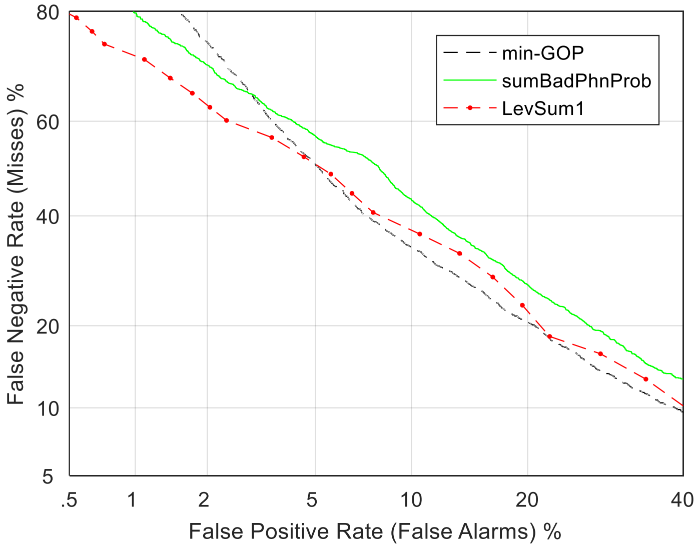


Figure 3.15: DET of min-GOP, sumBadPhnProb and LevSum1 as solo classification features, for SUB+PHO on CV-train, using automatic segmentation.

decoding systems over candidate segments (manual transcription), for correct words and for mispronunciations, as shown in Table 3.5, provides an interesting insight. As expected, the constrained phonetic lattice results in a low PER for correct words, since the sequence of correct pronunciation is much more probable. On the other hand, for mispronunciations, the

PER is higher using phone lattices since it has less freedom to recognize mispronounced phones. For the bigram, the higher PER on mispronunciations than correct words may reflect some problems of the manual transcription, as it is often hard to decide which sequence of phones was uttered in mispronunciations.

Table 3.5: Phone error rates (PER) of the two phone decoding approaches.

Phone decoding model	PER % correct word	PER % mispronunciations	PER % all
Bigram	22.81	36.03	24.12
Phonetic lattice	2.57	41.46	6.41

Returning to classification results, the DET curves of LevBigram1, LevPL1 and LevSum1 (their sum) are shown in Figure 3.16. As can be seen, LevPL1 performs worse at finding all mispronunciations, never going below 43% miss rate (with a threshold of distance 1, where the next lower one would be 0, for 100% false alarm and 0% miss). However, LevPL1 seems better for lower false alarms and the combination of both features clearly provides improved results.

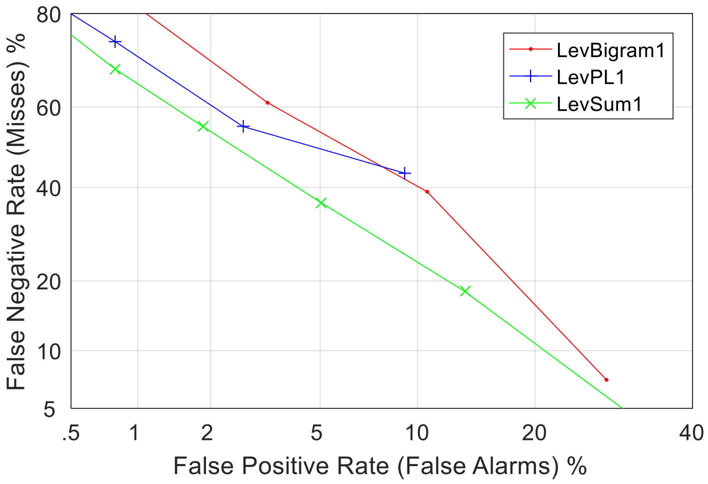


Figure 3.16: DET of Levenshtein edit distance features, for the classification of the SUB class, on the test set.

Comparing the use of manual segmentation versus automatic segmentation, the automatic one does result in slightly worse, albeit close, results, as shown in Figure 3.17. For a 5% false alarm rate, a 33.51% miss rate is obtained by LLR-spotter from manual segmentation, versus a 35.32% miss rate using automatic segmentation.

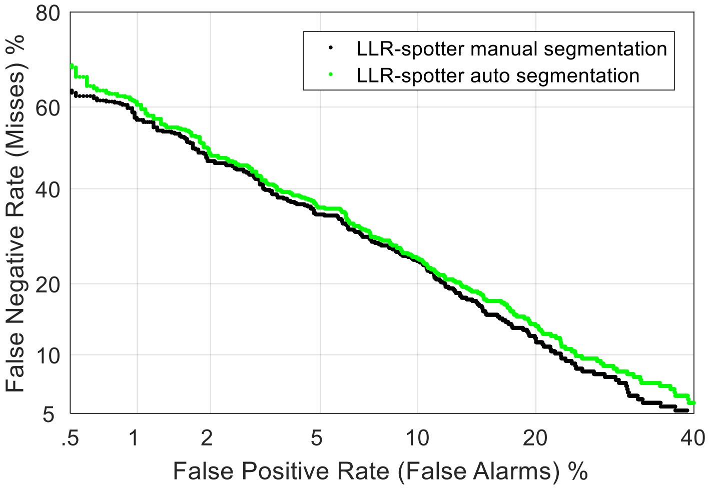


Figure 3.17: DET of LLR-spotter for the SUB+PHO classification on the test set by using manual or auto segmentations.

Table 3.6 summarizes the results of using the multi-feature models described in section 3.4.2, with the goal of combining the information of several features to improve classification. NN-step and SVM-step represent the use of the selected features by the best feature selection method for the same conditions (either Step-add or Step-remove). In addition to the cost obtained by the optimal thresholds from training, miss rates for the same 5% false alarm rate are indicated, although the operating points for the given costs vary slightly from 4% to 6% false alarm rate.

Table 3.6: Cost and miss rates at a 5% false alarm for the classification of SUB+PHO class vs. correct words, using multi-feature models (LLR-spotter included for comparison).

	CV-train				**Test**			
	Manual		**Auto**		**Manual**		**Auto**	
Classification Model	**Cost**	**Miss**	**Cost**	**Miss**	**Cost**	**Miss**	**Cost**	**Miss**
LLR-spotter	0.136	27.17	0.141	29.23	0.157	34.03	0.163	35.58
Logit-all	0.121	22.67	0.137	27.34	0.137	29.61	0.154	33.51
Step-add	0.120	22.35	0.136	27.26	0.141	29.61	0.150	33.25
Step-remove	0.121	22.59	0.136	27.22	0.140	29.87	0.153	33.51
NN	0.119	21.87	0.132	25.89	**0.136**	**27.79**	0.144	30.13
NN-step	**0.116**	**21.35**	0.133	26.17	0.140	28.83	0.144	32.47
SVM	0.117	22.03	**0.130**	**25.29**	0.139	29.35	**0.142**	**29.35**
SVM-step	0.118	21.63	0.130	25.53	0.139	30.65	0.152	33.77

A significant improvement was obtained by considering multiple features, and the best classifiers vary from neural networks to SVMs, with similar results. For the manual segmentation, neural networks provided better results, and using feature selection was greatly helpful for CV-train. For the other cases, stepwise feature selection was not helpful and the best results for automatic segmentation were obtained with SVMs. For the same 5% false alarm rates, the improvement of miss rate relative to the best individual feature (LLR-spotter) are 22% and 13% in CV-train (manual: 27.17% to 21.35%; auto: 29.23% to 25.29%) and 18% on the test set (manual: 34.03% to 27.79 %; auto: 35.58% to 29.35%).

Even if stepwise feature selection was only helpful for manual segmentation, analyzing which features are consistently selected may give an insight into the most relevant ones. For the Step-add feature selection, these are the features that are consistently selected for all the folds of cross-validation on the training set, for SUB+PHO:

- LLR-spotter;
- LLR-ali;
- mean-GOP;
- maxBadPhoneProb;
- LevDistPL1 or LevDistPL3;
- 1 combination of Levensthein distances - either LevSum3 or LevProd3;
- 1 normalization of LLR-spotter - LLR-s/Nchars, LLR-s/Nframes or LLR-s/Area;
- 1 interaction of LLR-spotter with phone lattice distance - LLR-s*LevPL1 or LLR-s*LevPL3.
- Area (LLR area from the spotting approach).

Most of the designed features prove to be relevant and apparently carry complementary information that enhances mispronunciation classification. Curiously, LLR-ali was always selected, even though it performs worse than LLR-spotter. It may be useful for cases where something extra is said at the beginning or end of words (for example, adding a plural suffix) and where by using the spotting approach on these segments, a correct pronunciation is found (LLR-spotter would hurt the classification), whereas the original segmentation encompassed the mispronunciation. Furthermore, mean-GOP and maxBadPhoneProb were chosen over min-GOP and sumBadPhoneProb. Effectively, even if min-GOP and sumBadPhoneProb are better individually (in minimum cost and in the stepwise criterion of deviance), after the first step when LLR-spotter is added to the stepwise model, mean-GOP and maxBadPhoneProb would provide better results if added. This is due to these two selected features having less correlation or sharing less information with LLR-spotter and other important features, and helping the model with alternative information.

To further analyze the improvement of using multi-feature models over LLR-spotter, Figure 3.18 and Figure 3.19 show the DET curves comparing the use of the individual LLR-

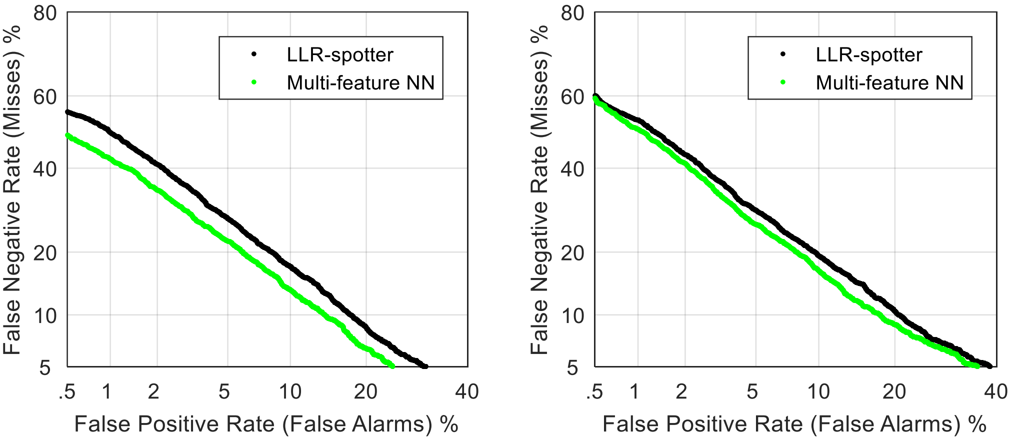


Figure 3.18: DET curves of LLR-spotter and multi-feature neural network using manual (left) and automatic (right) segmentations, for SUB+PHO CV-train classification.

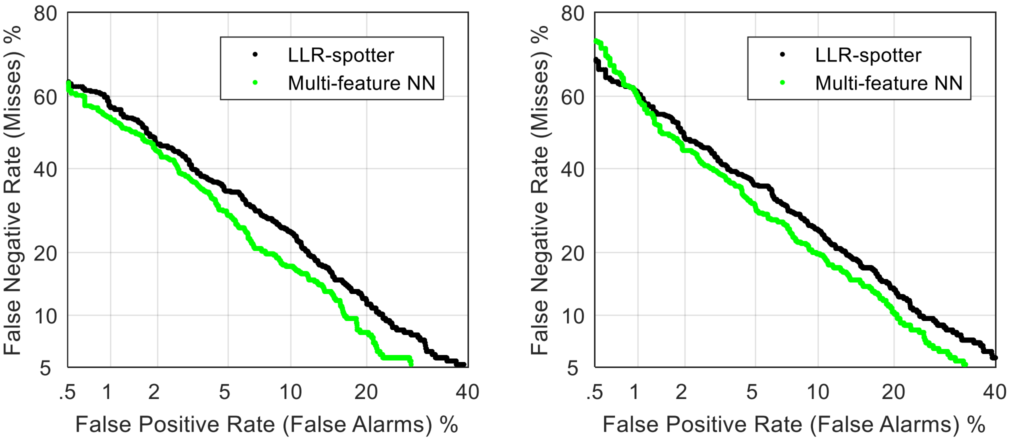


Figure 3.19: DET curves of LLR-spotter and multi-feature neural network using manual (left) and automatic (right) segmentations, for SUB+PHO test classification.

spotter feature versus a multi-feature model, in this case, the neural network using all features. It is apparent that the improvement from multiple features to only one feature from using manual annotation is larger than using automatic annotation, although there was still an improvement for the latter.

The results analyzed so far are for the broad class of mispronunciations (SUB+PHO). However, small mispronunciations (PHO) are especially difficult to detect, due to slight changes in one phoneme. Lowering this challenging aspect, Table 3.7 shows the results of classification of only the severe mispronunciation class (SUB) vs. correct words. Clear improvements were obtained for all stages, although, there was less improvement than other cases for CV-train using automatic segmentation. For the test set, automatic segmentation approaches the performance of manual segmentation. Results were still very similar for neural networks and SVMs. The same observations are shown by the DET curves of multi-feature models vs. LLR-spotter, depicted in Figure 3.20 and Figure 3.21 for CV-train and test, respectively.

Table 3.7: Cost and miss rates at a 5% false alarm for the classification of SUB class vs. correct words, using multi-feature models (LLR-spotter included for comparison).

	CV-train				**Test**			
	Manual		**Auto**		**Manual**		**Auto**	
Classification Model	**Cost**	**Miss**	**Cost**	**Miss**	**Cost**	**Miss**	**Cost**	**Miss**
LLR-spotter	0.113	21.01	0.121	23.08	0.114	21.43	0.119	23.38
Logit-all	0.098	15.96	0.117	21.27	0.092	16.88	0.107	18.83
Step-add	0.098	15.81	0.121	22.78	0.090	16.23	0.106	19.48
Step-remove	0.099	16.27	0.118	21.49	0.093	16.88	0.106	19.48
NN	0.095	15.06	0.112	20.06	0.091	13.64	0.090	13.64
NN-step	0.094	14.76	0.113	20.21	0.091	14.29	0.093	14.29
SVM	0.094	15.06	**0.110**	**19.31**	**0.088**	**13.64**	**0.089**	**13.64**
SVM-step	**0.094**	**14.46**	0.112	19.76	0.088	13.64	0.095	14.94

Overall, be it only LLR-spotter or a multi-feature model, it can be observed from the edges of the DET curves that there are cases of mispronunciation that are hardly detected, only with very high false alarm rates. There are also cases of correct word pronunciations that easily result in false alarms. Most of these, where the manual annotator did not indicate mispronunciation, were found to be due to two factors: noise simultaneous with speech and

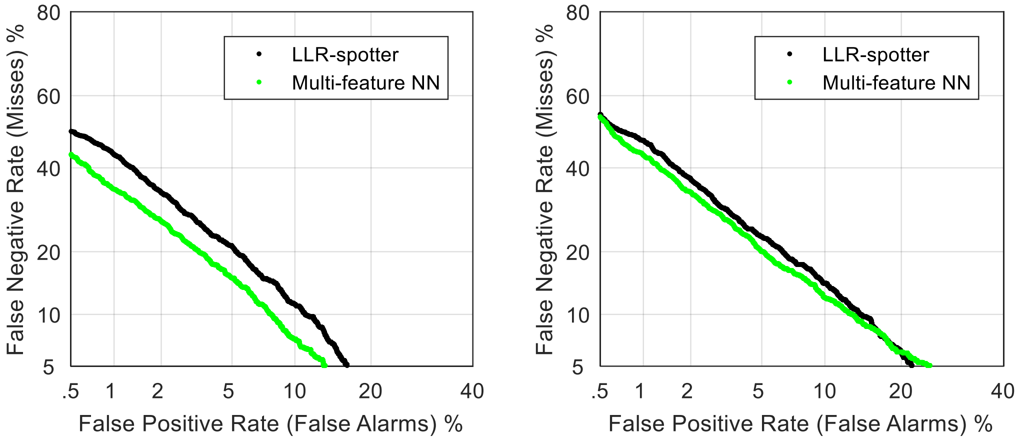


Figure 3.20: DET curves of LLR-spotter and multi-feature neural network using manual (left) and automatic (right) segmentations, for SUB CV-train classification.

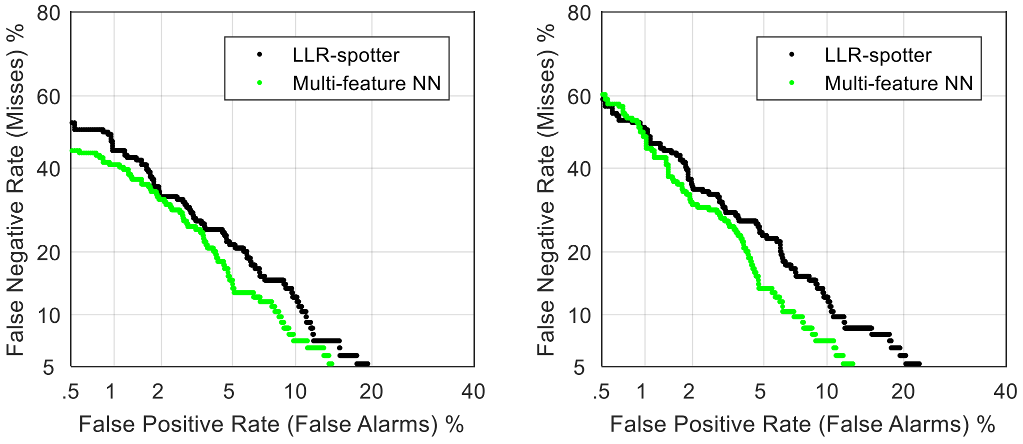


Figure 3.21: DET curves of LLR-spotter and multi-feature neural network using manual (left) and automatic (right) segmentations, for SUB test classification.

words with low vocal effort (whispering). Words with a low vocal effort often occur at the end of sentence with the final syllables of the word appearing unvoiced. It was attempted to add as features the word position in the sentence and a binary feature for being the last word, but they were never helpful.

There are two further main problems to tackle. The first is that the output of the phonetic recognizer is prone to errors, otherwise the match of the recognized phones to reference pronunciation would suffice. This was addressed by including several features that compensate for misrecognitions in some fashion (e.g., probability of mismatched phones and Levenshtein distance from a constrained phone decoding where the ideal sequence is highly probable). Nevertheless, by improving the accuracy of the phonetic recognizer, better results can be expected. The second problem is the subjective manual annotation of correct words and mispronunciations, where many cases are dubious for different manual annotators. Fixing annotator errors could have an effect on results but the methodology itself might not change. Not much can be done from an automatic perspective other than improving the reference by combining the opinions of multiple annotators.

The output of the methodology described in this chapter are automatic annotations of utterances of children reading, denoting extra content and correctly or incorrectly pronounced words. Extensions were marked if speech phones had a length higher than 0.60 seconds (empirically decided threshold). From this, several features can be extracted pertaining either to reading speed, disfluency rates or silence rates, which will be used to evaluate reading performance.

Chapter 4

Automatic Assessment of Reading Level

Although measuring correct words per minute can be an acceptable way to evaluate a child's reading, there may be other factors or specific problems that characterize the child's performance. Computing a score based on features from sentence reading tasks and pseudoword reading tasks can hopefully give an improved overall assessment of reading performance. Furthermore, there may other advantages to using a reading level: it can quickly provide a general overview of the literacy of the child and also allow that their evolution of reading proficiency be quickly observed from continued measurements. Previous works that focused on obtaining a reading performance or literacy level for children used individual word reading tasks as basis.

Duchateau et al. (2007) evaluated a child's reading ability by the number of correctly read words divided by time spent (same as correct words per minute) and show agreement to human evaluation with Cohen's Kappa (Cohen, 1960) above 0.6 when considering 5 performance classes.

Black et al. (2011) aimed to automatically provide a high-level literacy score. Eleven human evaluators of different backgrounds (linguistics, engineering, speech research) rated children's performance in individual word reading tasks with scores from 1 to 7. Using automatically extracted features and a selection of features based on pronunciation, fluency and speech rate, a Pearson correlation of 0.946 was achieved to predict mean evaluator's scores.

For the work on the LetsRead data, a reading performance score was proposed as a 0 to 5 level based loosely on grades 1 to 4. The general idea was that each grade would have an expected performance at the end of the school year and that there should be a score for lower than 1st grade level (zero) and a score for higher than 4th grade level (five). Therefore, the guidelines for the reading score were defined as:

- 0 – lower than expected at the end of the first grade.
- 1 – as expected at the end of the first grade.
- 2 – as expected at the end of the second grade.
- 3 – as expected at the end of the third grade.
- 4 – as expected at the end of the fourth grade.
- 5 – higher than expected at the end of the fourth grade.

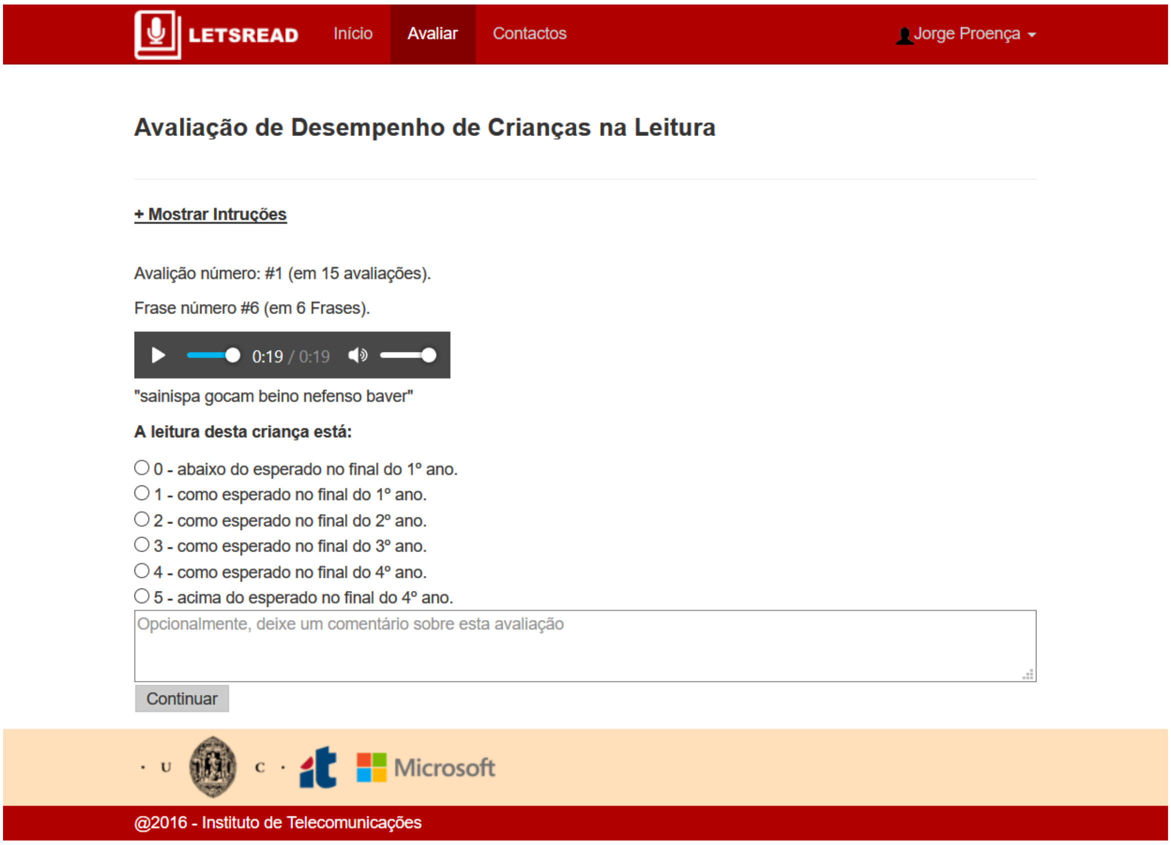
To analyze and automatically estimate the reading performance of the children of the LetsRead database, a ground truth of reading level is necessary. Although an assessment by speech or linguistic experts could be valid as done in Black et al. (2011), it was decided to get the opinion of elementary school teachers, since they are the ones that should accompany the children's evaluation and probably have more insight about reading level. Having this ground truth, regression models were trained based on several features extracted from child utterances, while comparing the use of manual and automatic annotations. Although the score guidelines of 0-5 are discrete, the average of teacher's opinions will be continuous values, as well as the predicted scores which will also be continuous.

4.1 Ground truth

In order to obtain a professional assessment for reading ability in children, elementary school teachers were asked to listen to utterances of reading tasks of children and provide a score for overall performance. This was achieved through a targeted crowdsourcing effort in Portugal, by closed divulgation through school groups ('agrupamentos escolares') of a website where teacher registration was needed[18]. A total of 151 people registered for evaluation, 109 completed the task, but only the first 100 to fully complete the task were

[18] http://lsi.co.it.pt/spl/letsread (only available in Portuguese)

considered to compile the ground truth. A sample screenshot of the website developed for this purpose is shown in Figure 4.1, during the evaluation stage.



Avaliação de Desempenho de Crianças na Leitura

+ Mostrar Intruções

Avalição número: #1 (em 15 avaliações).

Frase número #6 (em 6 Frases).

"sainispa gocam beino nefenso baver"

A leitura desta criança está:

○ 0 - abaixo do esperado no final do 1° ano.
○ 1 - como esperado no final do 1° ano.
○ 2 - como esperado no final do 2° ano.
○ 3 - como esperado no final do 3° ano.
○ 4 - como esperado no final do 4° ano.
○ 5 - acima do esperado no final do 4° ano.

Opcionalmente, deixe um comentário sobre esta avaliação

@2016 - Instituto de Telecomunicações

Figure 4.1: Screenshot of the website for child reading level evaluation by teachers, during an evaluation.

Each evaluator, after listening to 5 sentences and 5 pseudowords of a child, decided on a performance score between 0 and 5, based on the defined guidelines. These opinions will be used as a ground truth with which the automatically computed scores should be well correlated. A total of 150 children from the collected dataset were evaluated, 43 from the first grade, 40 from the second grade, 35 from the third grade and 32 from the fourth grade.

It was aimed that each evaluator should not spend more than 30 minutes on the requested task and it was necessary to balance this time limit with how many raters a child could be evaluated by. In the LetsRead corpus each child reads 20 sentences and 10 pseudowords. Initial tests for the rater effort showed that listening to only 5 sentences and 5 pseudowords is enough to provide an overall performance score, with the score rarely changing if more utterances are listened to. This assumption allowed to realistically aim for each child to be evaluated by at least 5 teachers. In the end, an average of 10 teachers evaluated each child.

Ten groups of evaluators evaluated different sets of 15 children (for a total of 150 children), with each group having an average of 10 evaluators (7 minimum, 13 maximum), for a total of 100 evaluators. The number varies as the collection process was affected by some allocated evaluators not finishing evaluations (not counted here). Therefore, although results may be shown for all evaluators, calculations of evaluator performance are done separately for each group. Each child was assessed by at least 7 evaluators. Figure 4.2 shows histograms per child of scores given by evaluators. Each column corresponds to a group of 15 children evaluated by the same group of evaluators.

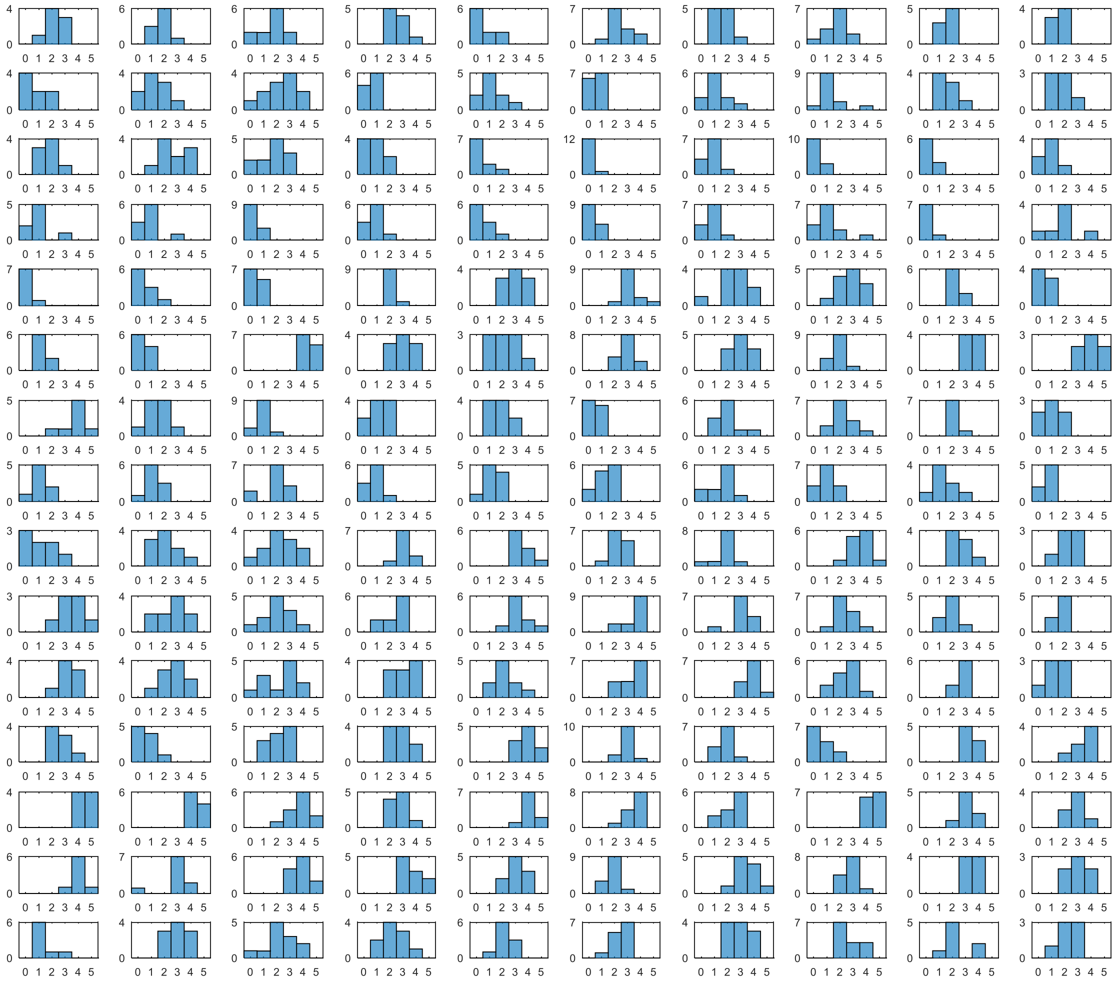


Figure 4.2: Histograms of reading performance score given by evaluators (x-axis) and evaluator count (y-axis) for 150 children (1 graph per child).

4.1.1 Evaluating evaluators

To measure agreement between evaluators, the Pearson's correlation of the 15 scores given by an evaluator to the 15 scores of another evaluator who assessed the same children can be computed, repeating for all evaluators of the same group. This pairwise correlation reflects the agreement between evaluators and can be used to identify 'bad' evaluators. For a group of 10 evaluators, there would be 9 values per evaluator. The mean of these 9 values for each evaluator describes their overall agreement with the group, shown in Figure 4.3.

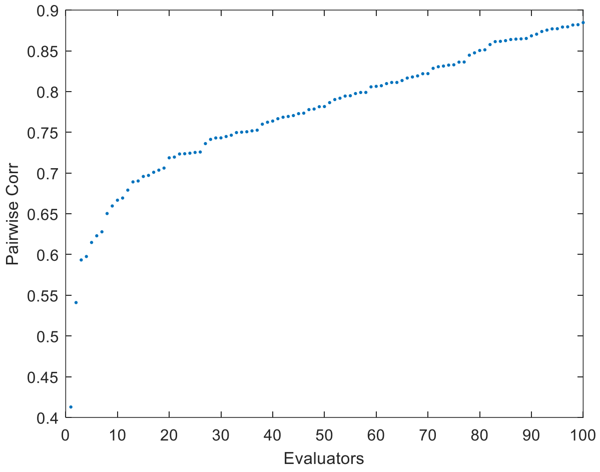


Figure 4.3: Mean pairwise correlations by evaluator, sorted from lowest to highest

There is one clear outlier with an average correlation of 0.413; this evaluator should be removed. It can be argued that the next 6 with low correlations also stand out, and the evaluators below 0.65 correlation should probably be removed for computing the ground truth. A problem is that some of them belong to the same group and by removing the worst of them, the others' average can improve and be higher than the threshold. Additionally, the pairwise correlations of all other evaluators must also be computed again. By removing the worst evaluators iteratively until none below 0.650 are kept, only 5 are eliminated.

As an alternative to pairwise correlation, the correlation of an evaluator's 15 scores with the 15 mean scores averaged from the other evaluators of the same group may be used (named as correlation to the mean). It gives higher values than the pairwise correlation but conclusions are similar. The final values for the two metrics are shown in Table 4.1.

Table 4.1: Final overall mean and standard deviation values of evaluator pairwise correlation and correlation to the mean of other evaluators.

Correlation	Mean ± S.D.	Maximum	Minimum
Pairwise	0.796 ± 0.060	0.885	0.657
To the mean	0.874 ± 0.069	0.967	0.679

4.1.2 Normalizing scores

As an alternative to using a mean score for a child from the raw values given by teachers, applying a z-normalization (z-norm) per evaluator, as in (4.1), can remove certain biases. These effects for an evaluator can be: i) constantly giving lower scores than the average ones; ii) constantly giving higher scores than the average ones; iii) constantly giving scores near the minimum and maximum; or iv) constantly giving scores near the middle.

$$x' = \frac{x-\mu}{\sigma} \qquad (4.1)$$

The z-norm for each evaluator changes their scores (x) by subtracting the mean of their 15 scores (μ) and dividing by the standard deviation of the 15 scores (σ). This results in values with zero mean and unitary standard deviation. Since these values do not fall in the intended scale of 0-5, they need to be reconstructed. This is done by multiplying by the overall standard deviation of all scores (1500) and adding the overall mean. This method is an alternative to scaling the minimum and maximum to 0 and 5 and can provide values slightly lower than 0 or higher than 5.

A pairwise correlation analysis would provide similar results to using non-normalized scores, since the changes are linear and correlation is linear. The same evaluators are removed. Figure 4.4 compares the final scores obtained using z-norm to the ones from a simple mean of raw scores, and small changes can be observed. The mean of differences are 0.062 ± 0.057, with a maximum difference of 0.329 (a score of 0.273 becoming -0.057 after normalization). The standard deviation of a child’s scores (the scores given by teachers for a child) lowers from an average 0.719 to 0.549 with normalization, showing that scores are more consistent than before. The mean of the transformed scores per child is the value taken as ground truth of reading performance score.

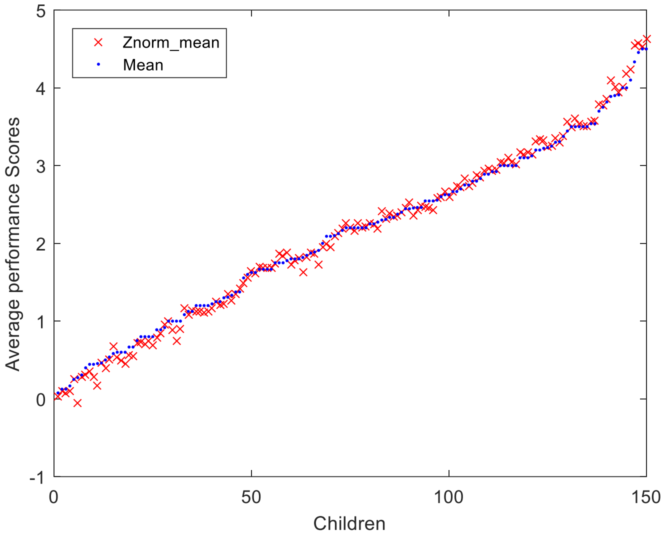


Figure 4.4: Mean of raw evaluator scores vs. mean of z-normalized scores.

4.2 Features

There should be several sources of information that teachers consider when deciding about overall reading aloud performance. Different types of features will be explored here, analyzing their performance individually in predicting performance score, and using a combination of them to train regression models. Two distinct sources for feature extraction will be considered: transcriptions with manual annotation (where all types of disfluencies are marked) and automatic transcriptions (where only mispronunciations, false starts, repetitions, intra-word pauses and extensions are marked). Although using manual annotation may provide the purest analysis of which features matter for reading performance evaluation, it is automatic feature annotation that is needed to build a practical system for performance evaluation.

The full set of considered features are described in Table 4.2. The same features are extracted from sentence reading tasks and pseudoword reading tasks separately, doubling the number of features shown. Features can be split into four groups:

- Reading speed features (1-6);
- Silence related features (7-9);
- Rates of disfluencies (10-22);

- Task-specific information (23-26).

Since there are a couple disfluency types that were not targeted in the automatic methods (deletions and insertions), features that depend on these disfluencies are only computed and analyzed for the manual annotation. From this point onwards, since features 1-26 repeat for sentences and pseudowords, sentence features will be addressed with the prefix ‘s’ and pseudoword features with the prefix ‘p’ (e.g., s1, s25, p1, p25).

Table 4.2: Enumeration of features for reading performance. These are extracted separately for sentence tasks and pseudoword tasks, leading to 52 features. Those with an asterisk are not computed for the automatic methods.

Group	Feat #	Abbreviation	Description
1	1	WPM	Words per minute (original prompt words over duration)
	2	CWPM	Correct Words Per Minute
	3	SyllsPM	Syllables per Minute (original prompt syllables)
	4	CSPM	Correct Syllables per Minute
	5	CharsPM	Characters per Minute (original prompt words’ characters)
	6	CCPM	Correct Characters per Minute
2	7	SILrate	Rate of Silence (Total Silence / Total time)
	8	SILini	Average Initial Silence time before first word
	9	SILiniRate	Initial Silence time / Total Time
3	10	SUBrate	Rate of SUB (number of SUB events / number of Words)
	11	PHOrate	Rate of PHO (number of PHO events / number of Words)
	12	PRErate	Rate of PRE (number of PRE events / number of Words)
	13	REPrate	Rate of REP (number of REP events / number of Words)
	14	PAUrate	Rate of PAU (number of PAU events / number of Words)
	15	DELrate *	Rate of DEL (number of DEL events / number of Words)
	16	EXTrate	Rate of EXT (number of EXT events / number of Words)
	17	INSrate *	Rate of INS (number of INS events / number of Words)
	18	MispR	Rate of SUB+PHO (Mispronunciations)
	19	ExtraR	Rate of PRE+REP (Extra segment disfluencies)
	20	SlowR	Rate of PAU+EXT (Slow reading disfluencies)
	21	FastR *	Rate of DEL+INS (Fast reading disfluencies)
	22	DisfR	Rate of Disfluencies (sum of all events / number of Words)
4	23	nSylls	Total number of syllables (original prompt syllables)
	24	nChars	Total number of characters (original prompt characters)
	25	Diff1	Difficulty 1 – Pronunciation rules without counting length
	26	Diff2	Difficulty 2 – Original difficulty index (rules and length)

An additional feature – Grade – that describes the school grade level each child is enrolled in (1 to 4) will also be considered. However, since a final application may not want to require knowledge of a child's grade level, models that do not use this feature should be built. Evaluators did not have grade information, although it is foreseeable that it will have some correlation to given scores, as average performance increases per grade (as previously observed for reading speed).

Some of the considered features are clearly very similar to each other, and some interesting conclusions may be drawn from analyzing their pairwise linear correlations. Given that this analysis results in a 53 by 53 symmetric matrix (which would be hard to present), these are the main observations for features obtained from manual annotation:

- Features s1-6 (reading speed) are highly correlated between each other (> 0.95).
- From p1-6, p5 is the most correlated with s1-6 (always ≈ 0.8), which is also the highest correlation between sentence and pseudoword features.
- Rate of silence in sentences (s7) is inversely correlated to reading speed s1-6 (≈ -0.69).
- Rate of disfluencies is significantly correlated with non-disfluency-related reading speed, e.g.: s1 and s22 with -0.71; p1 and p22 with -0.61.
- As expected, Grade is significantly correlated with task information s23-26 ($0.80 < \rho < 0.84$) and with reading speed: s1-6 ($0.67 < \rho < 0.72$); p1-6 ($0.45 < \rho < 0.52$).

Since some of these features provide similar information, it is foreseeable that between highly correlated features such as s1-6, one of them will be sufficient to predict reading performance. The next step in feature analysis will be to determine how each of them individually is able to predict ground truth scores.

4.3 Individual feature performance

The simplest way of fitting a feature to ground truth scores is by applying a linear transformation as in (4.2), trained by a linear regression (LR) model that minimizes the sum of squared errors (least squares).

$$\hat{y} = a^T x + b \tag{4.2}$$

In equation (4.2), $\hat{y}$ is the predicted output, x is the observation, a is the coefficient (weight) of the input and b is the intercept (bias) term. For a multi-feature regression, a is a vector of weights and x is a vector of observation values of multiple features. To evaluate the fit of predicted scores to the ground truth, two metrics will be considered: Pearson's correlation coefficient (ρ or Corr) and root mean squared error (RMSE), as described by equations (4.3) and (4.4). In both equations, $\hat{y}_i$ is the predicted output for a child, y_i is the reference score (ground truth given by the mean of normalized scores of teachers), $\mu_{\hat{y}}$ and μ_y are the mean scores for predicted outputs and ground truth and n is the number of children/scores analyzed ($n = 150$ in this case).

$$\rho = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \mu_{\hat{y}}\right)\left(y_i - \mu_y\right)}{\sqrt{\sum_{i=1}^{n}\left(\hat{y}_i - \mu_{\hat{y}}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \mu_y\right)^2}} \tag{4.3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n}} \tag{4.4}$$

To train and test regression models, a leave-one-out cross-validation with 150 folds will be considered, where 149 samples are used to train a model and 1 is left out for testing, until every sample is used in testing. The 150 resulting test values are aggregated from the different folds to compute Corr and RMSE. Although it is cumbersome to train 150 models, it is the best way to avoid dependence on different randomizations of folds that would lead to different average results.

Table 4.3 shows the performance of each feature if used individually to train a linear model. For comparison purposes, random performance leads to a correlation coefficient of 0 and RMSE of about 1.9. None of the strong correlation values are negative since any negatively correlated features are transformed with the linear model with a negative coefficient a. The correlation absolute value would be similar for the untransformed values, and features that have original negative correlations are silence rates (7-9) and disfluency rates (10-22).

Table 4.3: Performance of linear regression models predicting ground truth scores using individual features (average of leave-one-out cross-validation).

Sentences						Pseudowords				
		Manual		Auto			Manual		Auto	
Feat.	Abbr.	Corr	RMSE	Corr	RMSE	Feat.	Corr	RMSE	Corr	RMSE
s1	WPM	0.919	0.459	0.917	0.463	p1	0.744	0.776	0.744	0.776
s2	CWPM	0.928	0.434	0.923	0.447	p2	0.674	0.858	0.670	0.863
s3	SyllsPM	0.927	0.435	0.926	0.439	p3	0.764	0.750	0.760	0.755
s4	CSPM	0.938	0.402	0.930	0.429	p4	0.684	0.848	0.684	0.848
s5	CharsPM	0.931	0.424	0.930	0.428	p5	**0.805**	**0.689**	**0.803**	**0.693**
s6	CCPM	**0.940**	**0.397**	**0.931**	**0.425**	p6	0.703	0.827	0.691	0.840
s7	SILrate	0.647	0.885	0.736	0.787	p7	0.324	1.099	0.397	1.067
s8	SILini	0.347	1.091	0.480	1.019	p8	-0.157	1.176	-0.130	1.176
s9	SILiniRate	0.283	1.115	0.231	1.132	p9	-0.073	1.173	0.005	1.169
s10	SUBrate	0.376	1.105	0.615	0.916	p10	0.397	1.066	0.494	1.011
s11	PHOrate	0.394	1.073	N/A	N/A	p11	0.031	1.167	N/A	N/A
s12	PRErate	0.547	0.973	0.577	0.951	p12	0.131	1.156	0.217	1.135
s13	REPrate	0.190	1.142	0.378	1.076	p13	-0.008	1.170	0.010	1.170
s14	PAUrate	0.457	1.036	0.328	1.098	p14	0.124	1.156	-0.172	1.192
s15	DELrate	-0.037	1.172	N/A	N/A	p15	0.109	1.164	N/A	N/A
s16	EXTrate	0.350	1.092	0.381	1.073	p16	0.174	1.145	0.189	1.135
s17	INSrate	0.234	1.130	N/A	N/A	p17	-0.051	1.173	N/A	N/A
s18	MispR	0.525	0.991	0.615	0.916	p18	0.389	1.070	0.494	1.011
s19	ExtraR	0.498	1.008	0.555	0.968	p19	0.139	1.154	0.236	1.130
s20	SlowR	0.540	0.978	0.328	1.098	p20	0.247	1.127	-0.172	1.192
s21	FastR	0.171	1.146	N/A	N/A	p21	0.092	1.160	N/A	N/A
s22	DisfR	0.663	0.872	0.683	0.850	p22	0.490	1.013	0,530	0,985
s23	nSylls	0.456	1.034	0.456	1.034	p23	0.147	1.151	0.147	1.151
s24	nChars	0.483	1.018	0.483	1.018	p24	0.361	1.084	0.361	1.084
s25	Diff1	0.493	1.011	0.493	1.011	p25	0.464	1.029	0.464	1.029
s26	Diff2	0.491	1.012	0.491	1.012	p26	0.462	1.031	0.462	1.031

The best overall feature for both manual and automatic methods is s6: correct characters per minute in sentences (CCPM). A correlation of 0.94 for the manual feature indicates that this metric by itself can be a very good predictor of overall reading performance, proving that evaluators focus mostly on reading speed. Features based only on the number of disfluencies over the number of words (such as s22 – DisfR), although presenting a correlation around 0.7 for sentences, do not perform as well as reading speed, which shows that reading speed is of higher importance. Other reading speed metrics that do not depend on disfluencies (s1, s3 and s5) perform well but slightly worse than their counterparts using correctly read units. The opposite is observed for pseudoword features, with the non-disfluency-dependent reading speed features having significantly better performance, where the best one was p5: characters per minute of the original prompt (CharsPM). This may be due to poor performances in the pseudoword task, where values of 0 correct words per minute can be found, and the time it took to read them conveys more information than the number of correct readings. Figure 4.5 shows the relation between ground truth scores and the best features for sentences and pseudoword tasks, for the manual case, where there is evidence of a linear fit, especially for s6.

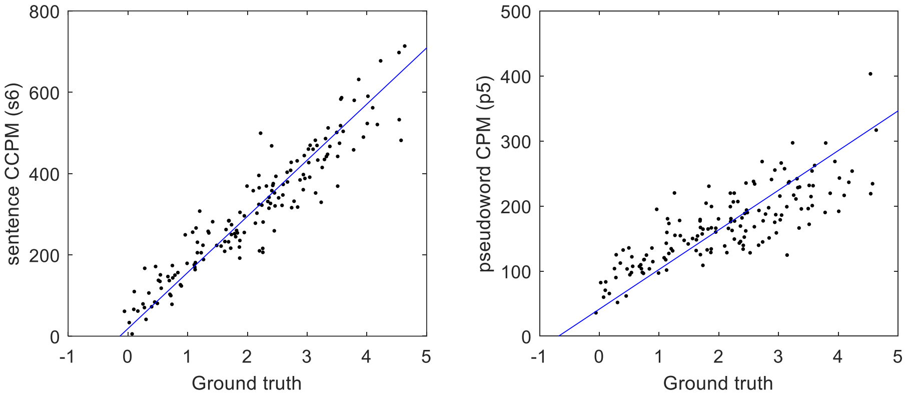


Figure 4.5: Ground truth scores vs. the best sentence feature (s6, left) and the best pseudoword feature (p5, right) from manual annotation features, with respective linear fits.

It must be emphasized that the results obtained with features 1, 3 and 5 (words, syllables and characters per minute) could be dependent on the conditions of the data. In the LetsRead dataset, it is typical that reading tasks are completed even if a lot of mistakes are made. It is very rare that a sentence is not finished and pseudoword lists are always completely attempted, which leads to these reading speed metrics, which only take into account the original prompt without considering if there are disfluencies, to have a certain significance for reading speed. If this were not the case, features 2, 4 and 6 (correct words, syllables and characters per minute) would be clearly preferable, since unfinished attempts or nonsense attempts would severely influence them. In a live application, these types of attempts should be expected. Additionally, there may be other cases where reading speed or correct words/characters per minute are not enough to characterize reading performance. For example, a very fast reader who often repeats words or gives a lot of false alarms but ends up pronouncing words correctly could have the same CWPM value as a reader with normal speed who reads without disfluencies. Even for incomplete attempts, CWPM could be of normal value, since it does not take into account the deleted words. For these cases, the features based on the relative number of disfluencies (e.g., s22 and p22) could be of help.

For the automatic features, the same conclusions apply, although s6 performs slightly worse than its manual equivalent. Unexpectedly, the performances of disfluency-dependent reading speed features (s2, s4 and s6) also fall closer to non-disfluency ones (s1, s3 and s5) probably due to certain disfluency detection errors, which may lead to the conclusion that it is not necessary to detect disfluencies. However, if the data presented more nonsense attempts, the results could be different.

Using the Grade feature for a linear regression model gives a positive weight, 0.647 Correlation and 0.886 RMSE, showing that scores increase per grade on average. Although the best feature – CCPM in sentences – already aggregates a lot of information (total reading time, length of tasks, length of words, correctly pronounced words), it is expected that combining the information of several of them will further approach the ratings by teachers.

4.4 Multi-feature Models

To combine multiple features in predictive models, several regression models are explored: multi-feature linear regression (LR), Gaussian process regression (GPR), and neural networks with one linear output. Several feature selection methods are also explored. One of the main problems to be tackled is overfitting, since a linear regression considering all the defined features will be strictly optimized for the training sets, with no regards for generalization. The selection of the most relevant features can be a way to minimize overfitting as well as using other regression methods such as GPR.

Gaussian process regression builds kernel-based probabilistic models to infer continuous values and is especially useful to avoid overfitting (Rasmussen and Williams, 2006). GPR brings together several concepts: Bayesian, kernelized, non-parametric, non-linear and modelling uncertainty. Since it is probabilistic, confidence intervals on a provided score are calculated, exemplified by Figure 4.6. GPR models were trained with a squared exponential kernel as the covariance function.

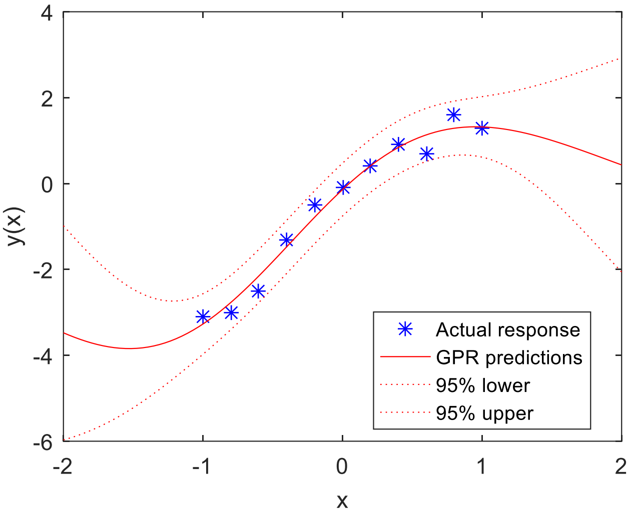


Figure 4.6: Theoretical example of Gaussian process regression (GPR) model with prediction confidence interval.

A simple neural network for regression was used with one hidden layer followed by a linear unit, providing linear outputs in the evaluation's range. Since the amount of data for training is low (149 samples maximum), the best results were obtained using only a single perceptron in the hidden layer. The networks were trained with Bayesian regularization to improve generalization (MacKay, 1992).

Stepwise regression can decide which features to include or remove iteratively for a regression model (Draper and Smith, 1998). Two stepwise approaches that start with no features included are explored: only adding features (add, forward or sequential) and bidirectional where features can be added and later removed (add+remove). The criterion to add a feature at each step is selecting the one that minimizes the sum of squared errors (SSE) when a linear regression is applied. However, a feature is only added if the decrease in SSE is statistically significant with a *p*-value of an *F*-statistic test lower than 0.05. Similarly, a feature can be removed at a certain step if its contribution to lowering SSE at this stage is not statistically significant. The trained linear regression is considered and also take the selected features to train GPR and NN models.

Another regression method that can be used for feature selection is the least absolute shrinkage and selection operator (LASSO), a regularization technique applied here for a regularized least-squares regression (Tibshirani, 1996). LASSO minimizes SSE but adds constraints to the sum of absolute values of coefficients of features, usually producing many weight coefficients equal to zero, usually for highly-correlated features. It produces a solution for a linear transformation but can also be a feature selection procedure (by selecting the features with weights different from zero), after which LR, GPR and NN can be applied.

Principal component analysis (PCA) is also explored to transform features into a set of linearly independent ones (Jolliffe, 2002). By applying this transformation to the entire set of features, it is hoped that the newly created features, especially the ones that explain most of the variance of the data, can be useful for regression. PCA is applied to the entire set of features and the ones that explain 95% of the variance are selected to train LR and GPR. Additionally, bidirectional stepwise regression is applied to the entire set of new features.

Using the stepwise algorithm for different cross-validation folds may result in different selections of features, with some being selected in several or all folds. With this knowledge, an analysis is done for results of pre-selecting the most common features given by the stepwise folds and then running LR or GPR using only those features. Although this feature selection depends on all the cross-validation folds using stepwise selection, the actual models trained (with a new leave one out cross-validation) do not depend on the left-out test values. These selections can be different for manual and automatic features. However, for manual features, the stepwise algorithm using leave-one-out folds selected the same four features 100% of the time, unlike with automatic features, where variations did occur. So, for the manual case, the following selected features are from a stepwise algorithm that allows the addition of features with an increased *p*-value for the *F*-test (0.15), now with variations occurring to which features are selected in each fold. The selections made are of features that were chosen in stepwise manner for at least a certain percentage of the folds: 80% (Sel80%), 40% (Sel40%) and 5% (Sel5%).

Table 4.4 summarizes the results of all the multi-feature models explored when considering manual or automatic transcriptions, with leave-one-out cross-validation. The first models are of the best individual feature (s6 – CCPM) and the selection of the best feature of sentences (s6) and the best of pseudowords (p5 – CPM). For manual features, NN proved to be the best model after a stepwise feature selection and the improvement to only using the best features is clearer in RMSE. Results were slightly worse with automatic features and, contrarily, the best results are without feature selection and the best model was GPR.

Table 4.4: Performance of multi-feature models on manual and automatic features (with test values after leave-one-out cross-validation).

Manual Features			Automatic Features		
Model	**Corr**	**RMSE**	**Model**	**Corr**	**RMSE**
LR (s6)	0.940	0.397	LR (s6)	0.931	0.425
LR (s6,p5)	0.943	0.386	LR (s6,p5)	0.933	0.419
GPR (s6,p5)	0.948	0.371	GPR (s6,p5)	0.938	0.403
LR all	0.926	0.442	LR all	0.931	0.426
GPR all	0.947	0.375	GPR all	**0.943**	**0.388**
NN all	0.938	0.403	NN all	0.939	0.399
Stepwise add + LR	0.947	0.373	Stepwise add + LR	0.919	0.458
Stepwise add+remove + LR	0.947	0.373	Stepwise add+remove + LR	0.919	0.458
Stepwise add + GPR	0.949	0.367	Stepwise add + GPR	0.932	0.422
Stepwise add + NN	**0.952**	**0.357**	Stepwise add + NN	0.925	0.442
LASSO	0.944	0.387	LASSO	0.932	0.423
LASSO + LR	0.942	0.392	LASSO + LR	0.932	0.421
LASSO + GPR	0.942	0.392	LASSO + GPR	0.939	0.400
LASSO + NN	0.948	0.357	LASSO + NN	0.935	0.411
PCA 95% + LR	0.917	0.465	PCA 95% + LR	0.916	0.467
PCA 95% + GPR	0.931	0.423	PCA 95% + GPR	0.927	0.436
PCA all + Stepwise + LR	0.909	0.488	PCA all + Stepwise + LR	0.938	0.404
PCA all + Stepwise + GPR	0.939	0.401	PCA all + Stepwise + GPR	0.936	0.408
LR Sel80% (s6,21;p1,25)	0.947	0.373	LR Sel80% (s3,6)	0.937	0.407
LR Sel40% (s6,21;p1,2,25)	0.947	0.373	LR Sel40% (s3,6,21;p22,25)	0.940	0.398
LR Sel5% (s1,6,18,21;p1,2,5,6,25)	0.946	0.377	LR Sel5% (s3,4,5,6,21;p1,4,19,22,25)	0.937	0.405
GPR Sel80% (s6,21;p1,25)	0.949	0.367	GPR Sel80% (s3,6)	0.940	0.397
GPR Sel40% (s6,21;p1,2,25)	**0.949**	**0.366**	GPR Sel40% (s3,6;p22,25)	0.940	0.398
GPR Sel5% (s1,6,18,21;p1,2,5,6,25)	0.947	0.373	GPR Sel5% (s3,4,5,6;p1,4,19,22,25)	**0.944**	**0.384**

The most notable observation is that GPR proved to be superior to LR at every stage, demonstrating its generalization capabilities. Successful feature selections only improved results slightly compared to using GPR over all features, which is already very robust to noise. The improvement of using multiple features instead of only s6 (CCPM) stands out more clearly in RMSE, going from 0.397 to 0.366 with manual features and from 0.425 to 0.384 with automatic features. Although using manual features provided the best overall results, the best automatic features model only presents a relative 0.5% lower correlation and 5% higher RMSE than the best manual model.

No features were removed during bidirectional stepwise regression. For LASSO, although it shows good performance, it never provided the best results. Applying PCA to the entire feature matrix provided worse results, although it shows better performance than stepwise selection over raw features in the automatic methods.

Analyzing the features commonly chosen by stepwise regression, the >40% selection provided slightly better results for manual annotation, with the selected features being:

- For all: CCPM (s6) and p25 (Diff1 – difficulty based on pronunciation rules only);
- For manual annotation: FastR (s21 – rate of deletions+insertions) and pseudo WPM (p1) and pseudo CWPM (p2);
- For automatic annotation: Sent SyllsPM (s3) and pseudo DisfR (p22).

This shows that reading speed of both sentences and pseudowords was relevant, as well as the difficulty of pseudowords based on dubious and infrequent pronunciation rules. The combined rate of deletions and insertions was also chosen for the manual annotation, with a negative weight, meaning that although these disfluencies are more common in higher grades, they are often given by fast speakers and this term might appear as a regulatory term to lower their scores that would otherwise be high. For the automatically obtained features selected from stepwise regression, reading speed of pseudowords was not chosen very often, although p22 (DisfR – rate of all disfluencies) does appear in the >40% selection. Nevertheless, the best model was obtained from the features appearing in more than 5% of the folds, which includes: reading speeds of both sentences and pseudowords (s3-6, p1 and p4), p19 (ExtraR – rate of false-starts+repetitions), p22 (DisfR – rate of all disfluencies) and p25 (Diff1 – difficulty based on pronunciation rules only). There are three common features with the manual analysis (s6, p1 and p25) with the rates of disfluencies being the additions that stand out. Since the automatic annotation does not detect deletions and insertions, feature s21 (ExtraR) could not be selected for the automatic analysis.

Figure 4.7 shows the predicted scores of the best model for automatic features – GPR Sel5% – with their corresponding Ground truth scores, as well as a 95% confidence interval given by the probabilistic GPR model. It can be seen that the GPR model fits most of the reference scores inside its 95% confidence interval, excluding some outliers.

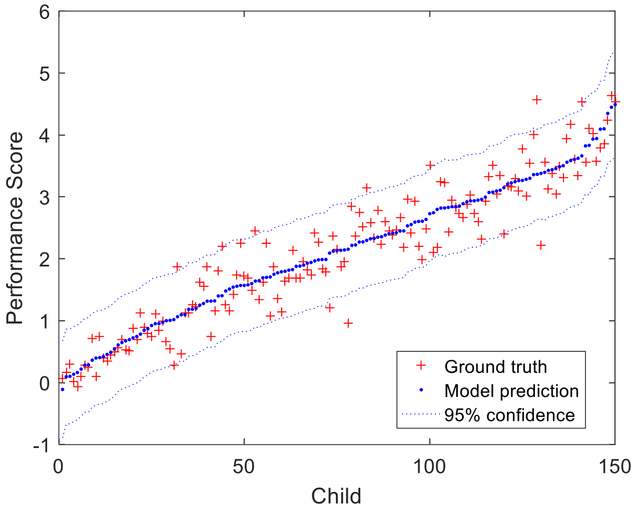


Figure 4.7: Per child reference scores (Ground truth) and predicted scores by the best performing Gaussian Process Regression (GPR) model using automatic features, including a 95% confidence interval of the GPR.

Figure 4.8 displays the reference ground truth scores with the predictions of the best model for automatic features (same as above), including the standard deviation (std) of the scores given by teachers for each child (the mean of those values results in the reference score for that child). This deviation, with an average of 0.549, reflects the evaluator uncertainty associated with scoring each child. The RMSE of the predicted scores by the model (0.384) is lower than evaluator standard deviation, and most predictions fall inside the deviation interval (again, with some outliers).

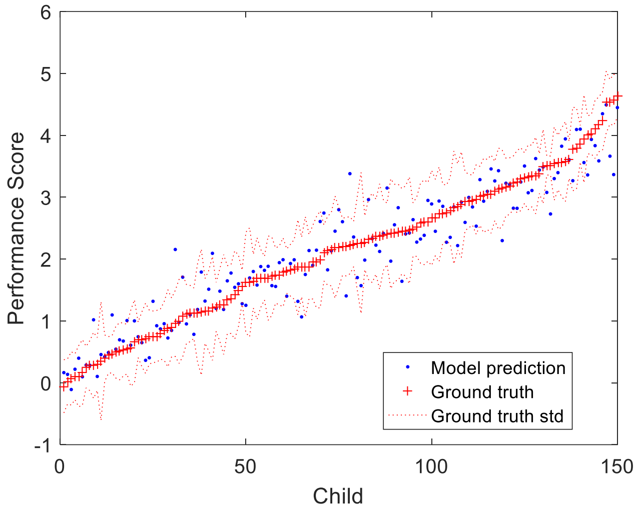


Figure 4.8: Per child predicted scores by the best-performing Gaussian process regression (GPR) model using automatic features and reference scores (Ground truth) including the standard deviation (std) interval of the opinion of teachers for each child.

Overall, since metrics based on both sentence reading and pseudowords reading tasks were used by the best performing models, it may be concluded that teachers gave their overall impression based on both tasks. Although reading speed features were the most important factors for reading ability assessment, detecting disfluencies proves to be of relevance as well, even for rates of specific types of disfluencies, which are different features than when they are considered to compute correct words/syllables/characters per minute. The difficulty of the tasks given is also an important factor to take into account when predicting reading performance and is possibly a normalizing factor. It is reflected in the selection of the difficulty of the pseudowords for both manual and automatic feature models and, indirectly, in the increase in performance obtained when using correct characters per minute instead of correct words per minute, since the length of words may be a measure of difficulty.

4.5 Additional results

The results reported in the previous subsections were published in (Proença et al., 2017b, 2017c). The automatic transcription used to extract features resulted from the decoding lattice without back-transitions described in section 3.3.1, which was not the best performing one for segmentation. Thus, results will now be analyzed when syllable-based decoding is used to get an automatic transcription as basis for features of reading performance.

Table 4.5 summarizes the results of using certain individual features to estimate reading performance. There is an improvement of the newest automatic annotation over the previous one for reading speed metrics and, surprisingly, s6 and p5 achieved practically the same performance as manual annotation.

Table 4.5: Correlation results of linear regression models predicting ground truth scores using individual features (average of leave-one-out cross-validation). AutoNew uses automatic segmentation from a syllable decoding strategy.

Feat.	Abbr.	Manual	Auto	AutoNew	Feat.	Manual	Auto	AutoNew
s1	WPM	0.919	0.917	0.917	p1	0.744	0.744	0.745
s2	CWPM	0.928	0.923	0.928	p2	0.674	0.670	0.668
s3	SyllsPM	0.927	0.926	0.926	p3	0.764	0.760	0.761
s4	CSPM	0.938	0.930	0.936	p4	0.684	0.684	0.675
s5	CharsPM	0.931	0.930	0.930	p5	0.805	0.803	**0.805**
s6	CCPM	0.940	0.931	**0.939**	p6	0.703	0.691	0.685
s18	MispR	0.525	0.615	0.562	p18	0.389	0.494	0.438
s19	ExtraR	0.498	0.555	0.539	p19	0.139	0.236	0.227
s20	SlowR	0.540	0.328	0.459	p20	0.247	-0.172	0.062
s21	FastR	0.171	N/A	N/A	p21	0.092	N/A	N/A
s22	DisfR	0.663	0.683	0.649	p22	0.490	0.530	0.500

Using multiple features in regression models, Table 4.7 summarizes the obtained results. Although the best automatic models did not get the same performance as manual annotation, there was still a slight improvement over previous automatic annotation attempts. Notably, stepwise feature selection models were more successful, although the best results were not obtained for them as was the case of manual annotation. Without feature selection, results were very close to or better than manual annotation for the considered models.

Table 4.7: Performance of multi-feature models on manual and automatic features (with test values after leave-one-out cross-validation).

	Manual		**Auto**		**AutoNew**	
Model	**Corr**	**RMSE**	**Corr**	**RMSE**	**Corr**	**RMSE**
LR (s6)	0.940	0.397	0.931	0.425	0.939	0.398
LR (s6,p5)	0.943	0.386	0.933	0.419	0.941	0.393
LR all	0.926	0.442	0.931	0.426	0.935	0.413
GPR all	0.947	0.375	**0.943**	**0.388**	**0.946**	**0.378**
NN all	0.938	0.403	0.939	0.399	0.940	0.398
Stepwise add + LR	0.947	0.373	0.919	0.458	0.938	0.404
Stepwise add + GPR	0.949	0.367	0.932	0.422	0.943	0.386
Stepwise add + NN	**0.952**	**0.357**	0.925	0.442	0.941	0.393

There are also interesting individual features that were not included initially: phones per minute (PhonesPM), correct phones per minute (CPhonesPM) and OLD20. As expected, the features based on phones have a very similar behavior as the ones based on characters/graphemes, although they perform slightly worse for pseudowords. OLD20 has worse performance than the other difficulty metrics, especially for pseudowords. When redoing multi-feature models, phones per minute is now often additionally selected during stepwise feature selection. The best result (Gaussian process regression using all features) improved from 0.946 to 0.947 correlation and from 0.378 to 0.374 RMSE, with the inclusion of these features. These results are presented in Table 4.6.

Table 4.6: Performance of individual features and multi-feature Gaussian process regression from automatic segmentation using a syllable-based strategy (with test values after leave-one-out cross-validation).

Task	Feat.	Corr	RMSE
Sentences	PhonesPM	0.930	0.426
	CPhonesPM	0.939	0.400
	OLD20	0.483	1.018
Pseudowords	PhonesPM	0.748	0.770
	CPhonesPM	0.684	0.847
	OLD20	0.251	1.125
All	GPR	**0.947**	**0.374**

4.6 Live Applications

There were two main applications of the developed work at the time of writing:

- A demo website to browse LetsRead data and check automatic live results of annotation and performance score.
- A prototype website that teachers can use to give reading tasks to students and check their performance results.

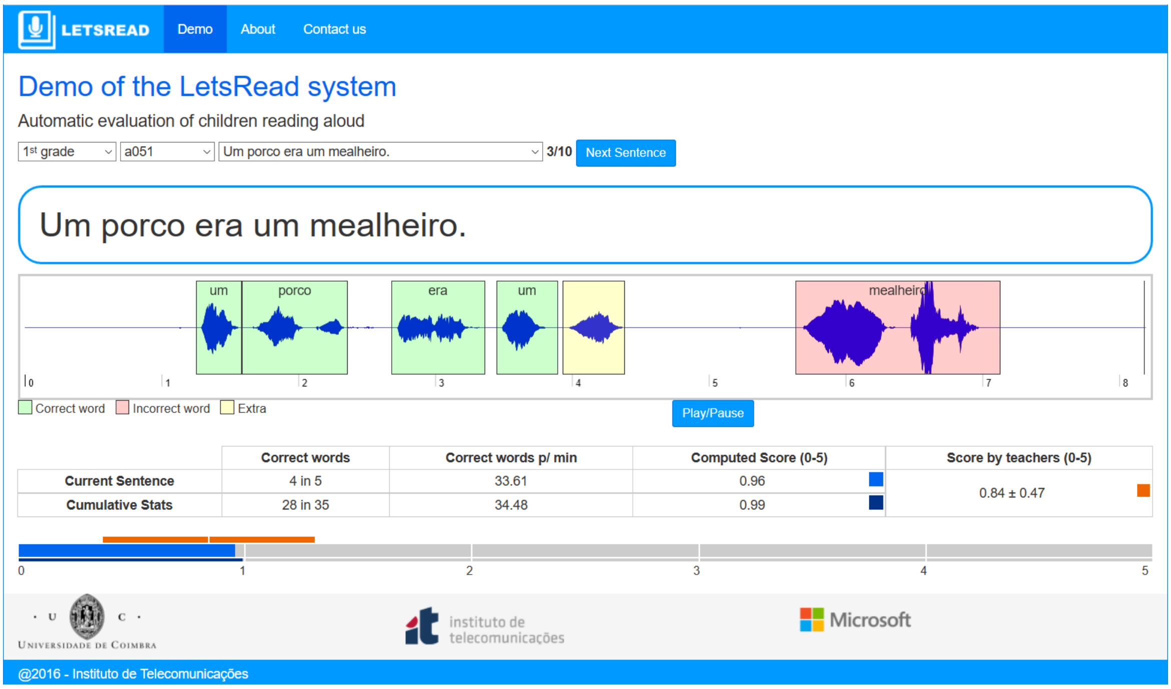
4.6.1 Demo

The demo application[19], reported in Proença et al. (2016e, 2016c), quickly allows to check the results of the proposed methods on children from the Letsread database. The audio is pre-recorded but, in the current iteration, the results are not pre-computed. A simulation of live audio input is done at the server level and utterances are automatically annotated, along with a calculation of performance metrics and reading score. A snapshot of the demo is presented in Figure 4.9. Specifically, for the demo:

- From the Ground truth by teachers of overall reading performance, the same utterances of 150 children can be selected and heard.
- The current sentence is shown in a large font, simulating an application where a child would have to read it live.
- Along with the audio signal, an automatic annotation is processed and presented where extra content and problematic words are detected.
- Statistics of correct words and overall score are computed per sentence and accumulated and re-computed as more utterances of that child are heard.
- The mean of overall performance score given by teachers (0-5) and standard deviations are presented.

[19] The LetsRead demo: http://hades.co.it.pt:9000/index_en.html – A short video with an overview of the project and of the demo can be found at: http://lsi.co.it.pt/spl/letsread/Letsread_demo.mp4

- Finally, the computed and accumulated scores are graphically displayed and compared to the ground truth.



Demo of the LetsRead system

Automatic evaluation of children reading aloud

1st grade | a051 | Um porco era um mealheiro. | 3/10 | Next Sentence

Um porco era um mealheiro.

🟩 Correct word 🟥 Incorrect word 🟨 Extra

	Correct words	Correct words p/ min	Computed Score (0-5)	Score by teachers (0-5)
Current Sentence	4 in 5	33.61	0.96 🟦	0.84 ± 0.47 🟧
Cumulative Stats	28 in 35	34.48	0.99 🟦	

@2016 - Instituto de Telecomunicações

Figure 4.9: Snapshot of the LetsRead demo application.

Node.js (Tilkov and Vinoski, 2010) is used for the server-side environment, with the audio processing and decoding being implemented in a C++ addon. This implementation has certain differences to the reported methodology, since Kaldi was not used. Only the long temporal context neural network that outputs phone posteriors is used, both as input for segmentation (syllable-based decoding) and for word likelihood estimation. This includes an implementation of temporal patterns (TRAPS) and of a Viterbi decoder. The grammars used for segmentation are the syllable-based lattices, automatically built for a given prompt. Taking as input the audio and prompt text of an utterance, the audio is parameterized and passed through the network where posterior probabilities are obtained. Then, syllable-based decoding is applied over the posteriors to get word segmentation/candidates and false-starts. For each candidate, its allowable pronunciations and co-articulations according to neighboring context are obtained and a log-likelihood ratio through a spotting approach (LLR-spotter) is computed for each candidate, considering all its allowable pronunciations.

Candidates are classified as mispronounced or not depending on a threshold of LLR-spotter values.

In the demo, after an utterance is selected by the web user, the server processes it and replies with a word-level segmentation with mispronunciation classification and marking extra content to be displayed along with the audio signal.

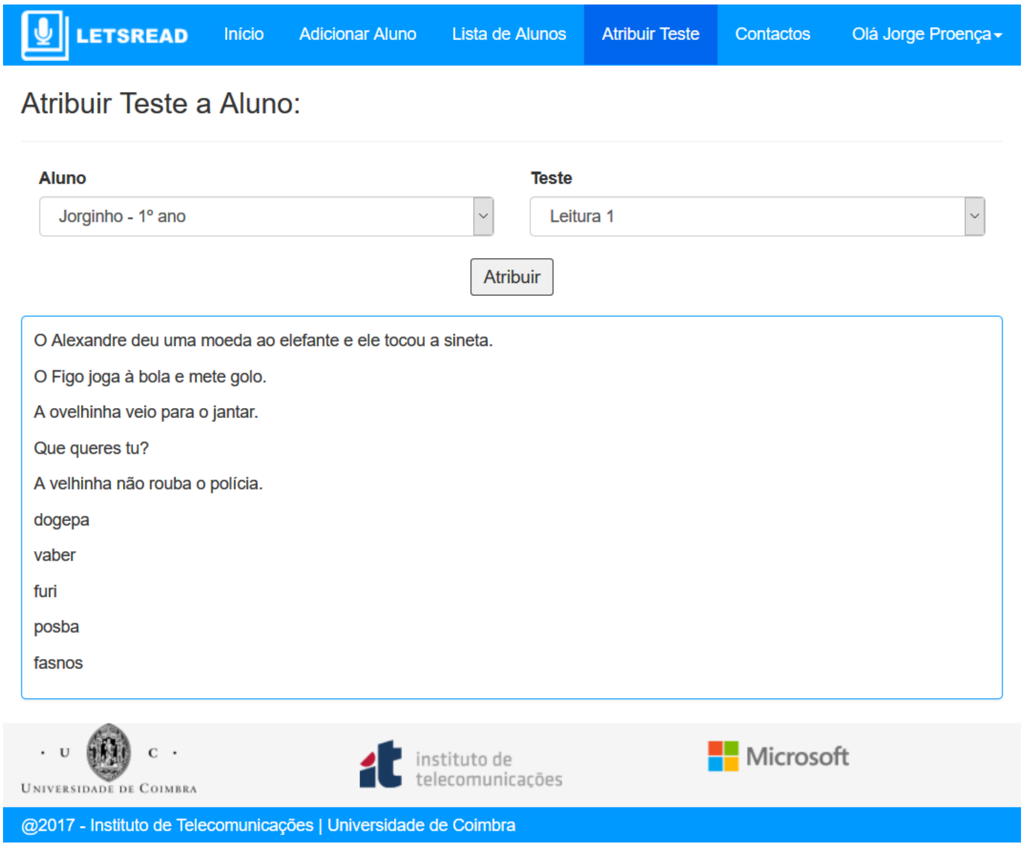
4.6.2 Prototype website for teachers

The prototype application[20] was a collaborative effort and was already the reported work of a master's student thesis (Almeida, 2018). The website is still a work in progress but it is the ideal application of the developed methodology. The platform is targeted for teachers, where each has a closed workspace where they mostly can:

- Add children to their list of students;
- Assign reading tasks to children, as depicted in Figure 4.10;
- Ask children to read the assigned task (requires a microphone input), as depicted in Figure 4.11. The audio for each utterance reading attempt is recorded and sent to the server, where annotation and performance are asynchronously computed;
- Visualize the results of all past reading tasks, with an interface similar to the LetsRead demo, where each utterance can be analyzed. The entire track record of the performance of reading tasks for the same child can be observed.

The development framework is similar to the demo application: node.js and the same C++ addon for audio processing and decoding. However, there are a lot more server-side details such as Express to serve dynamic webpages and database management. On the client-side during reading tasks, Web-audio API (Adenot et al., 2013) is used to extract microphone data, sending an utterance's full audio packet at the end of the utterance's reading attempt. No streaming is applied since getting the results for an utterance is not a time-sensitive issue (analysis of results can only done afterwards by the teacher). The computation of automatic annotation and reading performance is asynchronous.

[20] LetsRead prototype website: https://letsread.co.it.pt/ – The website is only available in Portuguese.



Atribuir Teste a Aluno:

Aluno: Jorginho - 1º ano

Teste: Leitura 1

O Alexandre deu uma moeda ao elefante e ele tocou a sineta.

O Figo joga à bola e mete golo.

A ovelhinha veio para o jantar.

Que queres tu?

A velhinha não rouba o polícia.

dogepa

vaber

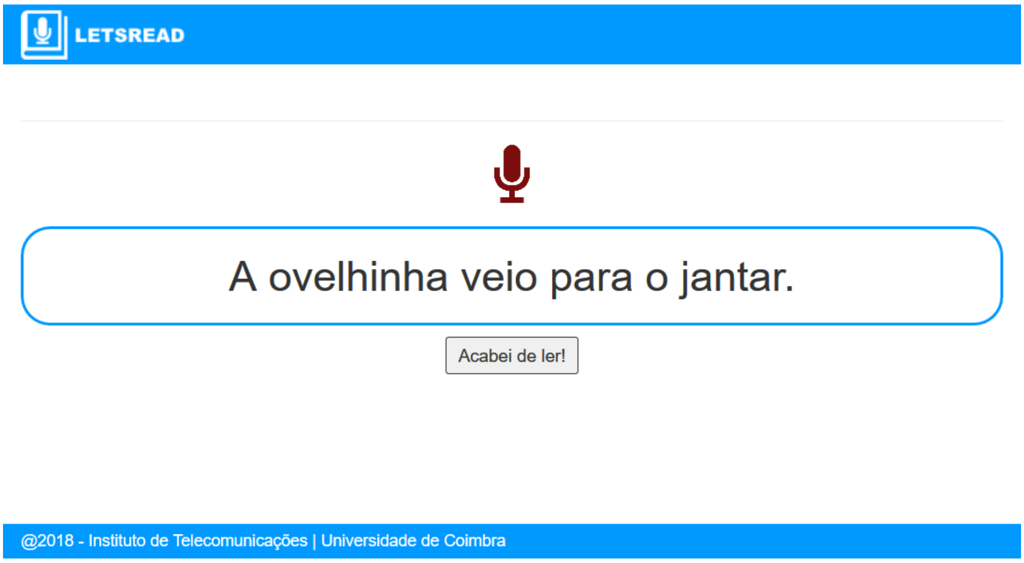
furi

posba

fasnos

@2017 - Instituto de Telecomunicações | Universidade de Coimbra

Figure 4.10: Snapshot of the LetsRead prototype website, during the assignment of a task.



A ovelhinha veio para o jantar.

@2018 - Instituto de Telecomunicações | Universidade de Coimbra

Figure 4.11: Snapshot of the LetsRead prototype website, during a reading task.

Since acoustic models were trained using only the LetsRead data, one of the main possible improvements is to make the application's acoustic models more robust to different acoustic conditions, which are now uncontrolled.

Furthermore, not all the methods explored for multiple feature combinations on either mispronunciation classification or reading performance score estimation were applied, due to time constraints. However, the best segmentation strategy of syllable-based decoding was applied. For mispronunciation classification, only the best feature – LRR-spotter – is used to decide if a word is correctly pronounced or not and, for reading score, a linear model of correct characters per minute of sentences and pseudowords is used. One could argue that the increase of real-time factor of processing multiple additional features and employing more complicated classification and regression models to achieve slightly improved results would not be worth it. However, this would probably not be an issue for the performance of the prototype application since results can only be consulted by the teacher after a child finishes a reading task, and not immediately after, so there is ample time for computation.

Chapter 5

Conclusions

5.1 Final remarks

The initial objectives proposed for this work were fulfilled, given than a solution was proposed and tested to automatically evaluate the reading ability of European Portuguese elementary school children. The final output is a prototype application that showcases the use of the developed methods.

A dataset was carefully designed and collected and several types of reading disfluencies were identified. The average performance of children was shown to increase with grade level on both sentence and pseudoword tasks, although there is a high variation within a grade level. Some of the most common disfluencies were targeted for automatic detection: false starts, repetitions, and mispronunciations. Some of the most common reading disfluencies were targeted for automatic detection: false starts, repetitions, and mispronunciations. Using task-specific lattices based on the prompt’s syllables, it was possible to detect almost 90% of disfluency events that result in extra segments, with a false alarm rate lower than 1%. Detecting mispronunciations proved to be much more challenging and, by using likelihood measures from the output of a neural network built towards phoneme recognition, in conjunction with phoneme sequence recognition features, more than 70% of mispronunciations were correctly identified with a 5% false alarm rate.

Regression models were trained to automatically predict reading performance scores and, although it is undeniable that correct words per minute read in sentences is already a very good measure for what teachers believe the reading level of a child to be, certain features were found to be effective in getting automatic scoring closer to the ground truth. Specifically, features relating to the performance in pseudoword tasks and the difficulty of these tasks were helpful when using either manually or automatically obtained features. Even if all disfluencies were not correctly identified with the automatic methods, the performance of models using features from the automatic annotation to predict overall reading score fell close to the performance based on manual annotation. Since metrics based on both sentence reading and pseudowords reading tasks were usually chosen during feature selection, it may be concluded that teachers gave their overall ratings based on both tasks.

Still, there are several limitations to be aware of. One is that no examples of reading individual words were collected, which are contemplated by curricular goals and are a separate task from the reading of sentences and pseudowords. Therefore, no direct comparisons were done with other works targeting reading level which used individual word reading.

Also, there was a large variety of different material in the tasks' prompts. This can be advantageous to build acoustic models and have sparse examples, but on the other hand proved to limit the consideration of specific reading error patterns and extrapolation of word difficulty. Therefore, recognition of mispronunciations was targeted in a general way through products of phonetic recognition. In fact, detecting mispronunciations proved to be the most challenging aspect of the developed work. The use of a log-likelihood ratio of ideal pronunciation versus free-phone-loop in a word-spotting fashion (LLR-spotter) proved successful in dealing with alignment problems of word candidate segmentations. If the trained phonetic recognizer were fully reliable, the recognition of the uttered phone sequence would suffice to realize if a word was mispronounced. However, that is not the case and most of the features proposed to be used in multi-feature models are included to both balance alignment problems and phonetic recognition imperfections.

Another limitation of the analyzed results is that only the LetsRead data was used to train and test acoustic models, decoding strategies, mispronunciation classification and reading level estimation. Since the data presents very similar acoustic conditions throughout, it is foreseeable that the trained acoustic models may not have the same performance for use-cases of different types of microphones, distances to microphone and acoustic environments. The data also has very consistent reading task attempts, where practically every prompt was read to the end (with varying success of pronunciation and taking a long amount of time if necessary). This lead to the fact that excluding the possibility of word deletions during segmentation presented better performance, which would not be the case if a lot of incomplete attempts were found. This may also indicate that the optimization of feature combinations (on all models) is highly adapted to the LetsRead data and not necessarily optimum for all applications.

There are also certain obvious aspects of a correct reading that were not automatically analyzed, especially those related to a correct prosody: the correct rhythm and pausing for a sentence's punctuation; and correct intonation. However, as previously stated, the obtained correlation of 0.95 to the opinion of teachers for reading level where only duration metrics relate to prosody, indicate that reading speed and correct reading speed are the most important aspects that a teacher considers to evaluate reading performance (correct prosody could, however, be directly correlated to higher reading speeds in the data). Focusing on reading speed and disfluency rates allowed the focus on using speech decoding and speech recognition paradigms to achieve the results.

As a final thought, these may be the three main challenges when trying to build a speech application for children: acoustic models, which need to be specifically tailored for children; different behaviors for speaking attempts or reading ability, which are reflected in numerous disfluencies or content extraneous to the task that an automatic system should consider; and the higher possibility of being uncooperative or not follow the application's instructions, which are probably interface issues. The LetsRead prototype does not really address the problem of maximizing the appeal and allure of the interface, since the reading task is usually done under adult supervision. But this should be a significant issue if a child's independence is targeted, especially the case for automatic reading tutors.

5.2 Future work

The predicted scores of overall reading performance fell mostly inside the standard deviation of human evaluation, although some outliers are found. Similarly, there are some ground truth scores that are outside the 95% confidence interval of the best GPR model. Further work needs to investigate which factors that were considered by teachers but not by the analyzed features lead to these outlier scores.

For technical enhancements, there are clearly matters to pursue: dealing with utterances with low vocal effort and improving phonetic recognition models. Speaker adaptation might also be useful at all stages, including adjusting phonetic recognition to individual reading speeds. The features considered for mispronunciation classification or reading score estimation may also have interactions (products, divisions) or transformations (squared, exponential, logarithm) which provide increased performance and should be investigated in more depth.

The results of this work should also be investigated on additional child data, not collected in an identical way as the LetsRead data. In fact, to achieve true robustness and readiness for a live application, additional child data from a variety of acoustic conditions should be acquired, preferably dozens or even hundreds of hours to be able to build deep neural networks for recognition. The web platform for teachers could be a source for new speech data. The platform would also benefit from applying all the best methods developed for word classification and score estimation using multiple features, with an increase of performance expected from the results of the investigated methods.

There is also the interesting possibility of exploring the use of the developed methodology on other languages, specifically English which has more interest for the research community and where more direct comparisons of achieved results can be done. Moreover, it could be interesting to analyze how well the techniques would behave for second language learning or older children. Applying the proposed methods to children older than 10 years would probably mean that new acoustic models would have to be built due to severe changes in the children’s voices. Additionally, the frequency of disfluent events (such as intra-word pauses) could be different, which would mean that the probabilities in models should be adjusted and that the features relevant for classification might even change. It is also a possibility to move from reading performance evaluation to reading tutoring, which would share several aspects of the methodology. The real-time factor of providing an automatic analysis of uttered speech would now become an important issue.

Annexes

Annex I. Reading task samples

Four reading tasks are presented on the following pages, one for each grade (1^{st}-4^{th}), used in the LetsRead database collection. Notice the use of larger sentences and harder vocabulary for the fourth grade exercise. On the final collection stages, five lists for each grade were iterated.

1st grade reading task (1 of 5):

- Pôs o fato de banho e foi para a praia correr.
- Os dias ficaram mais quentes e o patinho animou-se.
- A tia dá um funil azul ao Daniel e um anel à Joana.
- Para, escuta e olha.
- Com essa voz, acordavas-me a mim e aos meninos de noite!
- O soldado da farda azul passa pelo Rafael.
- Como esquecer o azul do mar?
- O gafanhoto também é conhecido por saltão.
- Correu para casa, porque se tinha esquecido das luvas.
- O patinho escondeu-se numa cabana onde morava uma velha.
- Como te chamas?
- Deixei o esconderijo e corri para a varanda do fundo.
- Não, não é contigo que vou subir ao altar.
- Não pode ser!
- Eu conheço um lobo com bom coração.
- A Fátima foi de mota até à vila de Fafe.
- A Lena chegou a casa, da escola, com sede e um pouco de fome.
- Seria possível?
- Eu tenho dois papás!
- Quem quer casar com a Carochinha, que é formosa e bonitinha?
- solsi
- ambafa
- mapos
- jusmes
- tutuvos
- tica
- feinocos
- vorapo
- jodas
- omendos

#####################################

2nd grade reading task (1 of 5):

- Ele ficou feliz e deu ao petiz um casaco com capuz.
- A maré estava muito baixa e a manhã estava linda.
- Sob a luz azul da lua, a cidade parecia suspensa no ar.
- O menino encontrou o vento do norte.
- Um galo andava à procura de bichinhos ou migalhas para comer.
- Pedro e Inês queriam-se bem, mas desencontraram-se durante a vida inteira.
- Queres saber?
- Nem todas as serpentes do mundo me podem fazer mal.
- -Porquê? - perguntaram os outros todos com grande espanto.
- O homem tocou-o com as esporas mas ele continuou imóvel e hirto.
- Sorriu de contente, olhou em volta e viu milhões de gotinhas como ela boiarem no mar.
- Domingo era o grande dia!
- Que mulher tão rabugenta!
- Certo dia, quando estava a varrer a cozinha, encontrou uma moeda de cinco réis.
- Há muito tempo, uma Carochinha, muito vaidosa, decidiu que queria casar.
- De facto, seguindo esses rastos, depressa chegou ao destino.
- A avó Rita ouvia rádio na rua e viu o rato.
- Era uma vez um gato maltês.
- Não estava habituada a meninos e aos seus doces passos de algodão.
- Como teria acontecido?
- sainispa
- gocam
- beino
- nefenso
- baver
- vesbo
- guira
- anor
- clapas
- condato

#####################################

3rd grade reading task (1 of 5):

- A Dinamarca fica no norte da Europa.
- O que é que tem trinta e dois moinhos a moer, uma vassoura a varrer e a pá a acartar?
- O João Ratão estava orgulhoso mas também muito nervoso.
- Parecia que as ondas iam cercar a casa e que o mar ia devorar o mundo.
- Os bancos do jardim, estrategicamente colocados sob a copa das árvores, estavam vazios.
- Lembra-me o senhor Joaquim, o meu amigo do Alentejo.
- Para além das intermináveis costas nuas e vazias, sem árvores, surgiram as primeiras palmeiras.
- Assim, amassou alguns pêssegos.
- Nos jardins há gladíolos e começa a estar mais calor.
- Trocaram juras de amor eterno e, no fim, choveu.
- O Sol parece mover-se no céu, do amanhecer até ao anoitecer, mas essa ilusão é causada pela rotação da Terra.
- O rato chegou-se ao velho pinheiro, e sem pedir licença, resolveu subir e apanhar pinhões.
- A Sara vai pintar um quadro.
- Era conhecida pela Casa do Ramalhete, ou simplesmente o Ramalhete.
- Pachorrenta como é de seu natural, a tartaruga abanou a cabeça, encolheu-se na casca, voltou a esticar-se.
- A quantidade de água potável disponível no nosso planeta é «uma gota no oceano».
- Ficou exausto, o meu soldado!
- Onde foste?
- A menina descobre o mundo, protegida por uma mãe que responde às suas dúvidas e que se deslumbra com as respostas.
- Gastei todas as minhas flechas, bati com a cabeça numa árvore e voltei ao palácio com um miserável ouriço.
- resder
- trinos
- rasol
- nicha
- pesvis
- fravis
- bragos
- oidos
- carto
- anasgar

#######################################

4th grade reading task (1 of 5):

- Um dia ele estatelou-se!
- Todas as luzes se apagaram à sua passagem e um manto muito grande, negro, de cetim, foi cobrindo a pouco e pouco o mundo inteiro.
- Diz o rato, que está no buraquinho, que está pronto para ser o padrinho.
- O Sol é uma enorme bola de gás quente, que se formou a partir de uma nuvem de gás e de pó que flutuava no espaço.
- Os guardas da fronteira prenderam os ladrões e levaram-nos à presença do rei.
- Vede que fresca fonte rega as flores, que lágrimas são a água e o nome Amores!
- O caçador sentiu que o ar lhe faltava, que as mãos lhe tremiam, e uma espécie de alegria muito grande.
- Eram três os Reis Magos.
- 'O romance da raposa' de Aquilino Ribeiro conta a história da Salta-Pocinhas.
- Oriana espreitou pela janela que não tinha vidro.
- Todos os meses, ela recebia uma longa carta, ou melhor, uma autêntica viagem dentro de um envelope selado.
- O rapaz estava condenado a carregar desde a nascença um nariz do tamanho de um chouriço e que, aos poucos, transforma a sua desgraça em graça.
- Ele nunca pôde voltar à sua cidade natal.
- Os mochos, as corujas e os guardas-noturnos queriam sair para o trabalho.
- Olhando para o seu antiquíssimo fato de trabalho, metade feito de estrelas, metade de escuros trapos, ela resolveu por uma vez ficar na cama.
- Ele contava histórias de lobos e ursos, e eu contava histórias de gnomos e anões.
- Até as árvores tinham agora uma nova roupagem.
- Mas isto é difícil!
- O pinheiro dava o néctar das suas flores às abelhas e a sua sombra às giestas que cresciam à sua volta.
- Ela conta que é irmã dos leões, que viveu na selva e que odeia ir ao jardim zoológico, onde estão presos todos os seus primos e familiares.
- tucha
- bolminar
- trinas
- hurascar
- solmi
- impencas
- deivinfes
- ulfes
- soche
- fentido

##

Annex II. Phonetic alphabet

Phonetic transcriptions in this document are shown in IPA. However, a subset of SAMPA is used on transcriptions of the LetsRead database. Table A.1 describes the used symbols with word examples. Stress marks are included before the stressed vowel, not before the syllable.

Table A.1: SAMPA and IPA phonetic alphabets for European Portuguese, with examples and special considerations for extensions and pauses.

	SAMPA	IPA	Example (orthographic, SAMPA, IPA)
1	a	a	amar, [6m"ar], [ɐmˈaɾ]
2	6	ɐ	camada, [k6m"ad6], [kɐmˈadɐ]
3	@	ə	semente, [s@m"e~t@], [səmˈẽtə]
4	e	e	letra, [l"etr6], [lˈetɾɐ]
5	E	ɛ	metro, [m"Etru], [mˈɛtɾu]
6	i	i	livru, [l"ivru], [lˈivɾu]
7	j	j	pai, [p"aj], [pˈaj]
8	o	o	como, [k"omu], [kˈomu]
9	O	ɔ	comes, [k"Om@S], [kˈɔməʃ]
10	u	u	comer, [kum"er], [kumˈeɾ]
11	w	w	pau, [p"aw], [pˈaw]
12	6~	ɐ̃	canto, [k"6~tu], [kˈɐ̃tu]
13	e~	ẽ	quente, [k"e~t@], [kˈẽtə]
14	i~	ĩ	cinto, [s"i~tu], [sˈĩtu]
15	o~	õ	conto, [k"o~tu], [kˈõtu]
16	u~	ũ	assunto, [6s"u~tu], [ɐsˈũtu]
17	j~	j̃	tem, [t"6~j~], [tˈɐ̃j̃]
18	w~	w̃	não, [n"6~w~], [nˈɐ̃w̃]
19	p	p	pato, [p"atu], [pˈatu]
20	t	t	sapato, [s6p"atu], [sɐpˈatu]
21	k	k	casa, [k"az6], [kˈazɐ]
22	b	b	barco, [b"arku], [bˈaɾku]
23	d	d	dado, [d"adu], [dˈadu]
24	g	g	galinha, [g6l"iJ6], [gɐlˈiɲɐ]
25	s	s	assim, [6s"i~], [ɐsˈĩ]
26	S	ʃ	achar, [6S"ar], [ɐʃˈaɾ]
27	f	f	faca, [f"ak6], [fˈakɐ]
28	z	z	zeloso, [z@l"ozu], [zəlˈozu]
29	Z	ʒ	jóia, [Z"Oj6], [ʒˈɔjɐ]
30	v	v	velocidade, [v@lusid"ad@], [vəlusidˈadə]
31	m	m	military, [milit"ar], [militˈaɾ]
32	n	n	banana, [b6n"6n6], [bɐnˈɐnɐ]

33	J	ɲ	banho, [b"6Ju], [bˈɐɲu]
34	l	l	lima, [l"im6], [lˈimɐ]
35	L	ʎ	talho, [t"aLu], [tˈaʎu]
36	r	ɾ	caro, [k"aru], [kˈaɾu]
37	R	ʀ	carro, [k"aRu], [kˈaʀu]
38	:	ː	vowel extension
39	...		pause

Annex III. Results of mispronunciation classification

Table A.2: Cost and miss rates at a 5% false alarm for the classification of SUB+PHO class vs. correct words, using individual features.

	CV-train				Test			
	Manual		Auto		Manual		Auto	
Feature	Cost	Miss	Cost	Miss	Cost	Miss	Cost	Miss
LLR-spotter	**0.136**	**27.17**	**0.141**	**29.23**	**0.157**	**34.03**	**0.163**	**35.58**
LLR-ali	0.171	38.32	0.191	45.04	0.187	44.94	0.192	48.05
min-GOP	**0.194**	48.23	**0.200**	**50.82**	**0.234**	**65.45**	**0.233**	**63.90**
mean-GOP	0.246	64.53	0.251	67.48	0.276	74.81	0.273	74.29
maxBadPhnProb	0.280	78.25	0.248	64.63	0.288	75.32	0.276	71.17
sumBadPhnProb	**0.232**	**56.62**	**0.231**	**56.88**	**0.241**	**60.00**	**0.247**	**63.12**
LevBigram1	0.248	65.29	0.247	61.98	0.263	71.43	0.262	71.69
LevBigram2	0.249	67.17	0.249	66.48	0.261	65.71	0.259	67.01
LevBigram3	0.248	64.85	0.247	64.71	0.257	65.19	0.258	66.23
LevPL1	0.227	54.65	0.235	57.33	0.242	66.75	0.252	70.91
LevPL2	0.227	56.70	0.235	59.78	0.243	67.01	0.253	58.96
LevPL3	0.227	57.70	0.235	60.98	0.242	66.75	0.252	62.34
LevSum1	0.199	47.87	0.206	52.51	0.212	50.39	0.221	53.77
LevSum2	0.201	49.48	0.209	52.63	0.214	51.43	0.222	54.29
LevSum3	0.199	48.19	**0.206**	**51.02**	**0.206**	**50.39**	**0.218**	**53.25**
LevProd1	0.196	48.35	0.211	54.15	0.224	56.10	0.228	55.58
LevProd2	0.199	48.35	0.211	54.15	0.227	57.14	0.230	58.18
LevProd3	**0.196**	**48.31**	0.210	54.11	0.225	56.36	0.229	57.92
LLR-s/Nphones	0.156	34.35	0.168	37.01	0.187	44.16	0.190	45.45
LLR-s*LevBigram1	0.159	35.11	0.162	36.25	0.179	40.52	0.185	42.60
LLR-s*LevPL1	0.179	48.68	0.194	53.71	0.203	52.73	0.208	56.36
Nframes	0.302	79.61	0.302	79.89	0.305	80.26	0.302	80.78
Area	0.319	87.84	0.319	88.32	0.315	88.05	0.314	88.57
Diff1	0.315	84.91	0.316	84.99	0.322	89.35	0.322	89.35
Diff2	0.313	83.55	0.314	83.70	0.322	87.27	0.322	87.27
OLD20	0.283	72.59	0.284	72.98	0.294	75.84	0.294	75.84
Nphones	0.300	78.81	0.301	78.92	0.298	73.51	0.299	73.77
Ngraph	0.309	80.06	0.309	80.13	0.307	77.14	0.307	77.14
Nsylls	0.316	84.11	0.316	84.87	0.312	76.88	0.312	76.88

References

Abdou, S.M., Hamid, S.E., Rashwan, M., Samir, A., Abdel-Hamid, O., Shahin, M., Nazih, W., 2006. Computer aided pronunciation learning system using speech recognition techniques, in: Proc. Interspeech 2006. Pittsburgh, USA, pp. 849–852.

Adenot, P., Wilson, C., Rogers, C., 2013. Web audio API. W3C 10.

Almeida, D., 2018. Implementação em tempo real de um sistema de avaliação automática de leitura de crianças (Masters Thesis). University of Coimbra.

Anguera, X., Rodriguez-Fuentes, L.J., Szöke, I., Buzo, A., Metze, F., 2014. Query by example search on speech at mediaeval 2014, in: Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain.

Black, M., Tepperman, J., Lee, S., Price, P., Narayanan, S., 2007. Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment, in: Proc. Interspeech 2007. Antwerp, Belgium, pp. 206–209.

Black, M.P., Tepperman, J., Narayanan, S.S., 2011. Automatic Prediction of Children's Reading Ability for High-Level Literacy Assessment. Trans. Audio, Speech and Lang. Proc. 19, 1015–1028. https://doi.org/10.1109/TASL.2010.2076389

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. https://doi.org/10.1145/1998384.1998390

Bose, A., Colangelo, A., Buchanan, L., 2011. Effect of phonetic complexity on word reading and repetition in deep dyslexia. Journal of Neurolinguistics 24, 435–444. https://doi.org/10.1016/j.jneuroling.2011.01.004

References

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

Candeias, S., Celorico, D., Proença, J., Veiga, A., Perdigão, F., 2013. HESITA(tions) in Portuguese: a database, in: ISCA, Interspeech Satellite Workshop on Disfluency in Spontaneous Speech - DiSS. KTH Royal Institute of Technology, Stockholm, Sweden, pp. 13–16.

Cincarek, T., Gruhn, R., Hacker, C., Nöth, E., Nakamura, S., 2009. Automatic pronunciation scoring of words and sentences independent from the non-native’s first language. Computer Speech & Language 23, 65–88. https://doi.org/10.1016/j.csl.2008.03.001

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20, 37–46. https://doi.org/10.1177/001316446002000104

Draper, N.R., Smith, H., 1998. Applied regression analysis, 3rd ed. Wiley.

Duchateau, J., Cleuren, L., hamme, H.V., Ghesquière, P., 2007. Automatic assessment of children’s reading level, in: Proc. Interspeech 2007. ISCA, Antwerp, Belgium, pp. 1210–1213.

Duchateau, J., Kong, Y.O., Cleuren, L., Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquière, P., Verhelst, W., hamme, H.V., 2009. Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. Speech Communication, Spoken Language Technology for Education Spoken Language 51, 985–994. https://doi.org/10.1016/j.specom.2009.04.010

Ferreira, A.J.S., 2007. Static features in real-time recognition of isolated vowels at high pitch. J. Acoust. Soc. Am. 122, 2389–2404. https://doi.org/10.1121/1.2772228

Fiscus, J.G., Ajot, J., Garofolo, J.S., Doddingtion, G., 2007. Results of the 2006 spoken term detection evaluation, in: Proc. SIGIR 2007. Amsterdam, Netherlands, pp. 51–57.

FIT, 2015. Phoneme recognizer based on long temporal context, Brno University of Technology [WWW Document]. URL http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context (accessed 5.6.15).

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.

Gray, S.S., Willett, D., Lu, J., Pinto, J., Maergner, P., Bodenstab, N., 2014. Child automatic speech recognition for US English: child interaction with living-room-electronic-devices., in: WOCCI. Singapore, pp. 21–26.

Hagan, M.T., Menhaj, M.B., 1994. Training feedforward networks with the Marquardt algorithm. IEEE transactions on Neural Networks 5, 989–993.

Hagen, A., Pellom, B., Cole, R., 2007. Highly Accurate Children's Speech Recognition for Interactive Reading Tutors Using Subword Units. Speech Communication 49, 861–873. https://doi.org/10.1016/j.specom.2007.05.004

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

Hämäläinen, A., Cho, H., Candeias, S., Pellegrini, T., Abad, A., Tjalve, M., Trancoso, I., Dias, M., 2014b. Automatically Recognising European Portuguese Children's Speech: Pronunciation Patterns Revealed by an Analysis of ASR Errors, in: Proc. International Conf. on Computational Processing of Portuguese - PROPOR. São Paulo, Brazil, pp. 1–11.

Hämäläinen, A., Rodrigues, S., Júdice, A., Silva, S.M., Calado, A., Pinto, F.M., Dias, M.S., 2013. The CNG Corpus of European Portuguese Children's Speech, in: Habernal, I., Matoušek, V. (Eds.), Text, Speech, and Dialogue, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 544–551.

Hasbrouck, J., Tindal, G.A., 2006. Oral reading fluency norms: A valuable assessment tool for reading teachers. The Reading Teacher 59, 636–644.

Jakielski, K.J., 1998. Motor organization in the acquisition of consonant clusters. PhD thesis, University of Texas at Austin, Ann Arbor Michigan: UMI Dissertation services.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer Science & Business Media.

Keuleers, E., Brysbaert, M., 2010. Wuggy: a multilingual pseudoword generator. Behav Res Methods 42, 627–633. https://doi.org/10.3758/BRM.42.3.627

Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. The Journal of the Acoustical Society of America 105, 1455–1468.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet Physics Doklady. pp. 707–710.

Li, X., Ju, Y.-C., Deng, L., Acero, A., 2007. Efficient and Robust Language Modeling in an Automatic Children's Reading Tutor System, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 193–196. https://doi.org/10.1109/ICASSP.2007.367196

Liao, H., Pundak, G., Siohan, O., Carroll, M.K., Coccaro, N., Jiang, Q.-M., Sainath, T.N., Senior, A., Beaufays, F., Bacchiani, M., 2015. Large vocabulary automatic speech recognition for children, in: Interspeech 2015. pp. 1611–1615.

Liu, Y., Shriberg, E., Stolcke, A., Harper, M.P., 2005. Comparing HMM, maximum entropy, and conditional random fields for disfluency detection, in: Proc. Interspeech. Citeseer, pp. 3313–3316.

Lopes, C., Veiga, A., Perdigão, F., 2012. A European Portuguese Children Speech Database for Computer Aided Speech Therapy, in: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Computational Processing of the Portuguese Language, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 368–374.

Lopes, J., Spear-Swerling, L., Oliveira, C., Velasquez, M., Almeida, L., Araújo, L., 2014. Ensino da leitura no 1º ciclo do ensino básico. Fundação Francisco Manuel dos Santos.

MacKay, D.J., 1992. A practical Bayesian framework for backpropagation networks. Neural computation 4, 448–472.

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E., 2009. PEAKS - A System for the Automatic Evaluation of Voice and Speech Disorders. Speech Communication 51, 425–437. https://doi.org/10.1016/j.specom.2009.01.004

Martin, A., R. Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of decision task performance, in: Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997. Greece.

Mateus, M.H.M., Pardal, E. d'Andrade, 2000. The Phonology of Portuguese. Oxford University Press.

Mc Laughlin, G.H., 1969. SMOG Grading-a New Readability Formula. Journal of Reading 12, 639–646.

Medeiros, H., Moniz, H., Batista, F., Trancoso, I., Nunes, L., others, 2013. Disfluency detection based on prosodic features for university lectures., in: Proc. Interspeech. Lyon, France, pp. 2629–2633.

Mendonça, G., Candeias, S., Perdigao, F., Shulby, C., Toniazzo, R., Klautau, A., Aluisio, S., 2014. A method for the extraction of phonetically-rich triphone sentences, in: Proc. of the International Telecommunications Symposium (ITS). São Paulo, Brazil, pp. 1–5. https://doi.org/10.1109/ITS.2014.6947957

Metze, F., Anguera, X., Barnard, E., Davel, M., Gravier, G., 2014. Language independent search in MediaEval's Spoken Web Search task. Computer Speech & Language 28, 1066–1082. https://doi.org/10.1016/j.csl.2013.12.004

Moniz, H., Batista, F., Mata, A.I., Trancoso, I., 2014. Speaking style effects in the production of disfluencies. Speech Communication 65, 20–35. https://doi.org/10.1016/j.specom.2014.05.004

Mostow, J., Roth, S.F., Hauptmann, A.G., Kane, M., 1994. A Prototype Reading Coach That Listens, in: Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1), AAAI '94. American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 785–792.

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, USA.

Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., Robert, M., 2014. The Goodness of Pronunciation algorithm applied to disordered speech. Presented at the The 15th Annual Conference of the International Speech Communication Association - INTERSPEECH 2014, International Speech Communication Association (ISCA), pp. 1463–1467.

Pellegrini, T., Hämäläinen, A., de Mareüil, P.B., Tjalve, M., Trancoso, I., Candeias, S., Dias, M.S., Braga, D., 2013. A corpus-based study of elderly and young speakers of European Portuguese: acoustic correlates and their impact on speech recognition performance., in: Proc. Interspeech 2013. Lyon, France, pp. 852–856.

Potamianos, A., Narayanan, S., 2003. Robust recognition of children’s speech. IEEE Transactions on Speech and Audio Processing 11, 603–616. https://doi.org/10.1109/TSA.2003.818026

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi Speech Recognition Toolkit, in: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, Hilton Waikoloa Village, Big Island, Hawaii, US.

Proença, J., Castela, L., Perdigão, F., 2015a. The SPL-IT-UC Query by Example Search on Speech system for MediaEval 2015, in: Working Notes Proceedings of the MediaEval 2015 Workshop. Wurzen, Germany.

Proença, J., Celorico, D., Candeias, S., Lopes, C., Perdigão, F., 2016a. The LetsRead Corpus of Portuguese Children Reading Aloud for Performance Evaluation, in: Proc of the 10th Edition of the Language Resources and Evaluation Conference (LREC 2016). Portorož, Slovenia.

Proença, J., Celorico, D., Candeias, S., Lopes, C., Perdigão, F., 2015b. Children’s Reading Aloud Performance: a Database and Automatic Detection of Disfluencies, in: Proc. Interspeech 2015. Dresden, Germany, pp. 1655–1659.

Proença, J., Celorico, D., Candeias, S., Lopes, C., Veiga, A., Perdigão, F., 2015c. A Database of Children Reading Aloud for Reading Performance Evaluation, in: Proc. of the 10th Conference on Telecommunications (Conftele). Aveiro, Portugal.

Proença, J., Celorico, D., Lopes, C., Candeias, S., Perdigão, F., 2016b. Automatic Annotation of Disfluent Speech in Children’s Reading Tasks, in: Proc. IX Jornadas En Tecnologías Del Habla and V Iberian SLTech Workshop - IberSPEECH’2016. Lisbon, Portugal, pp. 172–181.

Proença, J., Celorico, D., Lopes, C., Candeias, S., Perdigão, F., 2016c. LetsRead – Tool to Automatically Evaluate Children's Reading Aloud Performance. Presented at the International Conf. on Computational Processing of Portuguese - PROPOR, Demonstration session, Tomar, Portugal, pp. 15–17.

Proença, J., Celorico, D., Lopes, C., Sales Dias, M., Tjalve, M., Stolcke, A., Candeias, S., Perdigão, F., 2016d. Design and Analysis of a Database to Evaluate Children's Reading Aloud Performance, in: International Conf. on Computational Processing of Portuguese - PROPOR. Tomar, Portugal.

Proença, J., Costa, O.C., Celorico, D., Candeias, S., Perdigão, F., 2015d. Automatic Detection of Disfluencies in Children Reading Aloud Using Task Specific Lattices, in: Proc. of the 10th Conference on Telecommunications (Conftele). Aveiro, Portugal.

Proença, J., Lopes, C., Candeias, S., Perdigão, F., 2016e. LetsRead demo – Automatic Evaluation of Children's Reading Aloud Performance, in: Proc. IX Jornadas En Tecnologías Del Habla and V Iberian SLTech Workshop - IberSPEECH'2016. Lisbon, Portugal, pp. 433–435.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., Perdigão, F., 2018. Mispronunciation Detection in Children's Reading of Sentences. IEEE Transactions on Audio, Speech, and Language Processing (accepted).

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., Perdigão, F., 2017a. Detection of Mispronunciations and Disfluencies in Children Reading Aloud, in: Proc. Interspeech 2017. Stockholm, Sweden.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., Perdigão, F., 2017b. Automatic evaluation of reading aloud performance in children. Speech Communication 94, 1–14. https://doi.org/10.1016/j.specom.2017.08.006

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., Perdigão, F., 2017c. Automatic Evaluation of Children Reading Aloud on Sentences and Pseudowords, in: Proc. Interspeech 2017. Stockholm, Sweden.

Proença, J., Perdigão, F., 2016a. Segmented Dynamic Time Warping for Spoken Query-by-Example Search, in: Interspeech 2016. pp. 750–754. https://doi.org/10.21437/Interspeech.2016-1276

Proença, J., Perdigão, F., 2016b. The SPL-IT-UC QbESTD systems for Albayzin 2016 Search on Speech, in: Proc. IX Jornadas En Tecnologías Del Habla and V Iberian SLTech Workshop - IberSPEECH'2016. Lisbon, Portugal, pp. 43–50.

Proença, J., Veiga, A., Perdigão, F., 2015e. Query by Example Search with Segmented Dynamic Time Warping for Non-Exact Spoken Queries, in: Proc 23rd European Signal Processing Conference (EUSIPCO). Nice, France, pp. 1691–1695.

Proença, J., Veiga, A., Perdigão, F., 2014. The SPL-IT Query by Example Search on Speech system for MediaEval 2014, in: Working Notes Proceedings of the Mediaeval 2014 Workshop. Barcelona, Spain.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge, Massachusetts.

Rocha, P., Santos, D., 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa, in: Proc. PROPOR 2000. Atibaia, SP, Brazil, pp. 131–140.

Santos, A.L., Généreux, M., Cardoso, A., Agostinho, C., Abalada, S., 2014. A Corpus of European Portuguese Child and Child-directed Speech, in: Chair), N.C. (Conference, Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14). European Language Resources Association (ELRA), Reykjavik, Iceland.

Saz, O., Yin, S.-C., Lleida, E., Rose, R., Vaquero, C., Rodríguez, W.R., 2009. Tools and Technologies for Computer-Aided Speech and Language Therapy. Speech Communication, Spoken Language Technology for Education 51, 948–967. https://doi.org/10.1016/j.specom.2009.04.006

Shriberg, E.E., 1994. Preliminaries to a Theory of Speech Disfluencies. University of California, Berkeley.

Silva, A., Marques, C., Baptista, J., Jr, A.F., Mamede, N., 2012. REAP.PT Serious Games for Learning Portuguese, in: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Computational Processing of the Portuguese Language, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 248–259.

Soares, A.P., Medeiros, J.C., Simões, A., Machado, J., Costa, A., Iriarte, Á., de Almeida, J.J., Pinheiro, A.P., Comesaña, M., 2014. ESCOLEX: a grade-level lexical database from European Portuguese elementary to middle school textbooks. Behav Res Methods 46, 240–253. https://doi.org/10.3758/s13428-013-0350-1

Song, P., 2007. Correlated Data Analysis: Modeling, Analytics, and Applications, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-71393-9

Szoke, I., Rodriguez-Fuentes, L., Buzo, A., Anguera, X., Metze, F., Proença, J., Lojka, M., Xiong, X., 2015. Query by Example Search on Speech at Mediaeval 2015, in: Working Notes Proceedings of the MediaEval 2015 Workshop. Wurzen, Germany.

Tejedor, J., Toledano, D.T., Anguera, X., Varona, A., Hurtado, L.F., Miguel, A., Colás, J., 2013. Query-by-Example Spoken Term Detection ALBAYZIN 2012 evaluation: overview, systems, results, and discussion. EURASIP Journal on Audio, Speech, and Music Processing 2013, 23. https://doi.org/10.1186/1687-4722-2013-23

References

Tejedor, J., Toledano, D.T., Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., 2016. Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations. EURASIP Journal on Audio, Speech, and Music Processing 2016, 1. https://doi.org/10.1186/s13636-016-0080-2

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58, 267–288.

Tilkov, S., Vinoski, S., 2010. Node.js: Using JavaScript to Build High-Performance Network Programs. IEEE Internet Computing 14, 80–83. https://doi.org/10.1109/MIC.2010.145

Veiga, A., Celorico, D., Proença, J., Candeias, S., Perdigão, F., 2012. Prosodic and Phonetic Features for Speaking Styles Classification and Detection, in: Toledano, D.T., Giménez, A.O., Teixeira, A., Rodríguez, J.G., Gómez, L.H., Hernández, R.S.S., Castro, D.R. (Eds.), Advances in Speech and Language Technologies for Iberian Languages. Springer Berlin Heidelberg, pp. 89–98.

Veiga, A., Lopes, C., Sá, L., Perdigão, F., 2014. Acoustic Similarity Scores for Keyword Spotting, in: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Nunes, M. das G.V. (Eds.), Computational Processing of the Portuguese Language. Springer International Publishing, pp. 48–58.

Witt, S.M., Young, S.J., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication 30, 95–108. https://doi.org/10.1016/S0167-6393(99)00044-8

Yarkoni, T., Balota, D., Yap, M., 2008. Moving beyond Coltheart's N: A new measure of orthographic similarity. Psychonomic Bulletin & Review 15, 971–979. https://doi.org/10.3758/PBR.15.5.971

Yilmaz, E., Pelemans, J., hamme, H.V., 2014. Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model, in: Proc. Interspeech 2014. Singapore, pp. 969–972.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., 2006. The HTK book (v3. 4). Cambridge University.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., others, 1997. The HTK book. Entropic Cambridge Research Laboratory Cambridge.