



Prosodic Exercises for Children with ASD via Virtual Therapy

Mariana Sofia da Silva Sousa

Thesis to obtain the Master of Science Degree in
Electrical and Computer Engineering

Supervisor(s): Prof. Isabel Maria Martins Trancoso
Prof. Helena Gorete Silva Moniz

Examination Committee

Chairperson: Prof. João Fernando Cardoso Silva Sequeira
Supervisor: Prof. Isabel Maria Martins Trancoso
Member of the Committee: Prof. José Alberto Rodrigues Pereira Sardinha

June 2017

"The purpose of learning is growth, and our minds, unlike our bodies, can continue growing as long as we live." - Mortimer Adler

Acknowledgments

First of all, I would like to thank my supervisors, Professor Isabel Trancoso and Professor Helena Moniz, for all the guidance and knowledge transmitted, and also for the opportunity to achieve my goals throughout this journey.

I would like to thank all my colleagues from F2f that somehow helped and contributed to the concretization of this work, specially Professor Fernando Batista and Rubén Solera.

A special thank to INSIDE and RAGE, for giving me the opportunity to collaborate with them.

To all my friends who were always present, thank you for the encouragement and support.

To my parents, for always being there for me and making it possible for me to achieve my goals, and for all the support they gave me throughout this long and exhausting phase.

Last, but definitely not least, a special thanks to my boyfriend, Paulo, who supported me every single day along this journey. His constant encouragement, specially in the demotivating moments, gave me the strength to pursue and accomplish this work.

Resumo

As doenças do espectro do autismo, como o próprio nome indica, são doenças de espectro, o que significa que existe um elevado grau de variação na forma como este afeta as pessoas. Isto culmina num conjunto de inúmeras disfuncionalidades de desenvolvimento, que podem causar desafios a nível social, de comunicação e comportamentais, incluindo dificuldades na destreza da linguagem verbal. É sabido que, apesar do largo espectro, a caracterização da linguagem de crianças autistas tem sido consensual na literatura, como desprovida da riqueza de parâmetros prosódicos manifestados por crianças saudáveis, tais como aspetos emocionais que se refletem na interação comunicativa.

O uso de tecnologia como uma ferramenta de ensino tem vindo a crescer de dia para dia, e a apresentação de exercícios educacionais através de dispositivos eletrónicos, tal como o uso de terapias virtuais, robôs ou de aplicações móveis, revela-se mais atrativa e capacitiva para as crianças, quando comparada com os métodos tradicionais.

No presente projeto, foram desenvolvidos exercícios prosódicos, nomeadamente um método de avaliação de entoação através de um exercício de imitação, para desenvolvimento e enriquecimento das capacidades prosódicas de crianças com doenças do espectro do autismo, para auxílio de terapia. O método de avaliação de entoação foi avaliado, atingindo valores de precisão entre 70% e 83.3%, dependendo do conjunto de características adaptado (pitch, energia, MFCCs e informação proveniente de pseudo-sílabas), e também fazendo a fusão de todas as características. Apesar de, inicialmente, a intenção fosse a integração dos exercícios referidos numa plataforma existente para crianças diagnosticadas com doenças do espectro do autismo, a implementação atual consiste numa aplicação móvel.

Palavras-chave: Doenças do Espectro do Autismo, disfuncionalidades de desenvolvimento, parâmetro prosódicos, avaliação de entoação, aplicação móvel

Abstract

Autism Spectrum Disorder (ASD), as the name indicates, is a spectrum disorder, which means that there is a wide degree of variation in the way it affects people. It is known that, even though it has a huge spectrum, the characterization of the speech of autistic children has been consensual in the literature as devoid of wealth prosodic parameters manifested by healthy children, such as the emotional aspects that are reflected in communicative interaction.

The use of technology as a teaching tool has been growing and the presentation of educational exercises through electronic devices reveals itself as more attractive and captivating for children when compared with traditional methods.

In this project, we developed prosodic exercises, namely an intonation assessment method in an imitation task, where the main focus is the development and enrichment of prosodic abilities of children with autism spectrum disorders, as a complement to therapy sessions. We evaluated the intonation assessment method, achieving accuracy values between 70% and 83.3%, depending on the feature set adapted (pitch, power, Mel-frequency Cepstrum Coefficients (MFCCs), and pseudo-syllable information), and also by making a fusion of all features. Although the original intention was to integrate these exercises in an existing platform for children diagnosed with ASD, the current implementation is a stand-alone mobile application.

Keywords: Autism Spectrum Disorders, developmental disabilities, prosodic parameters, intonation assessment, mobile application

Contents

Acknowledgments	v
Resumo	vii
Abstract	ix
List of Tables	xv
List of Figures	xvii
Nomenclature	xix
1 Introduction	1
1.1 Motivation	1
1.2 Framework	3
1.2.1 RAGE	3
1.2.2 INSIDE	3
1.3 Goals	4
1.4 Document Outline	4
2 Background	7
2.1 Autism Spectrum Disorder	7
2.2 Emotions and Speech	9
2.3 Tests and Diagnosis	10
2.4 Therapies and Interventions	11
2.4.1 Applied Behaviour Analysis (ABA)	11
2.4.2 Floortime	13
2.4.3 Son-Rise	13
2.4.4 Relationship Development Intervention (RDI)	13
2.4.5 Training and Education of Autistic and Related Communication Handicapped Children (TEACCH)	14
2.4.6 Social Communication/Emotional Regulation/Transactional Support (SCERTS)	14
2.4.7 PEPS-C	15
2.4.8 PECS	15
2.5 Summary	15

3	Related Work	17
3.1	Technology for children with ASD	17
3.1.1	Characterization of user needs	17
3.1.2	Software for children with ASD	18
3.1.3	Discussion	24
3.2	Intonation Assessment	27
3.2.1	Pronunciation Evaluation	27
3.2.2	Nativeness, Fluency and Intonation Evaluation	28
3.2.3	Discussion	29
3.3	Summary	30
4	Intonation Assessment Method	31
4.0.1	Data Collection	31
4.1	Feature Extraction	32
4.1.1	Pitch	32
4.1.2	Power	33
4.1.3	MFCCs Extraction	34
4.1.4	Pseudo-syllables Extraction	35
4.2	Dynamic Time Warping	37
4.3	Classification	38
4.4	Tests and Results	39
4.5	Summary	41
5	New Contributions to the Virtual Therapist	43
5.1	Requirements Analysis and Definition	43
5.1.1	Functional Requirements	44
5.1.2	Non-Functional Requirements	44
5.2	Recorded Stimuli	44
5.3	Recorded Utterances	45
5.4	New Exercises	47
5.4.1	Intonation Distinction	48
5.4.2	Intonation Imitation	49
5.4.3	Affect Recognition	50
5.4.4	Up/Down Recognition	50
5.5	Summary	52
6	Conclusions and Future Work	53
6.1	Conclusions	53
6.2	Future Work	54
	Bibliography	55

A	Database	A.1
B	Recorded Utterances and Stimuli	B.2

List of Tables

3.1	Summary of Related Work	26
4.1	Stimuli database	32
4.2	Results obtained with MFCCs.	40
4.3	Results obtained with Pitch.	40
4.4	Results obtained with Power.	40
4.5	Results obtained with Pseudo-syllables features.	41
4.6	Final Results.	41
5.1	Recorded stimuli for the intonation distinction task.	45
5.2	Recorded stimuli for the affection recognition task.	45
5.3	Recorded stimuli for the imitation task.	45
A.1	Complete database.	A.1
B.1	Recorded Stimuli.	B.3

List of Figures

1.1	Estimated point prevalence of ASDs for male and female in 2010.	2
2.1	Autism Symptoms.	11
2.2	PEPS-C Structure.	15
2.3	Example of PECS.	16
3.1	VITHEA-KIDS platform.	20
3.2	Learning with Rufus.	22
3.3	Emotions and Feelings - Autism.	23
3.4	Screenshots of SPEAKall application.	23
3.5	iCommunicate application.	24
3.6	TalkInPicture application.	24
3.7	Diagram of detection of sentence stress.	29
4.1	Intonation Assessment Method.	31
4.2	Pitch extracted using <i>Aubio</i> library.	33
4.3	Pitch contour represented in two different scales.	33
4.4	Power representation.	34
4.5	How to extract MFCCs features.	35
4.6	Visualization of MFCCs series.	35
4.7	Window that allow us to modify some values.	36
4.8	TextGrid made by the script.	36
4.9	Example of info window produced by the <i>Praat</i> script.	36
4.10	DTW graphics representation.	37
4.11	Representation of the optimal warping path.	38
4.12	Cost between a stimulus and its imitation.	39
4.13	Set of informations given by the train class.	39
4.14	Testing the method.	39
5.1	Implemented Structure.	47
5.2	Architecture of the equal/different concept exercise.	48
5.3	Layout of the equal/different concept exercise.	48

5.4	Architecture and layout of the intonation distinction exercise.	49
5.5	Architecture of the intonation imitation exercise.	49
5.6	Layout of the intonation imitation exercise.	50
5.7	Architecture of the affect recognition exercise.	50
5.8	Layout of the affect recognition exercise.	51
5.9	Architecture of the high/low recognition exercise.	51
5.10	Layout of the high/low recognition exercise.	52

Acronyms

AAC	Argumentative and Alternative Communication
ABA	Applied Behaviour Analysis
ASD	Autism Spectrum Disorder
DSM	Diagnostic and Statistical Manual of Mental Disorders
DTW	Dynamic Time Warping
EP	European Portuguese
ESDM	Early Start Denver Model
FCT	Fundação para a Ciência e Tecnologia
INSIDE	Intelligent Networked Robot Systems for Symbiotic Interaction with Children with Impaired Development
MFC	Mel-frequency Cepstrum
MFCCs	Mel-frequency Cepstrum Coefficients
PDD	Pervasive Developmental Disorders
PDD-NOS	Pervasive Developmental Disorders - Not Otherwise Specified
PECS	Picture Exchange Communication System

PEPS-C	Profiling Elements of Prosody in Speech - Communication
PRT	Pivotal Response Treatment
RAGE	Realising and Applied Gaming Ecosystem
RDI	Relationship Development Intervention
SCERTS	Social Communication/Emotional Regulation/Transactional Support
TEACCH	Training and Education of Autistic and Related Communication Handicapped Children
VB	Verbal Behaviour
VITHEA	Virtual Therapist for Aphasia Treatment
VM	Video Modelling

Chapter 1

Introduction

The goal of this Master thesis is the development of an attractive mobile application that would help the acquisition of prosodic skills to children diagnosed with ASD. In this chapter we present the motivation for the development of this application, the goals we intend to achieve, as well as the outline for the remaining sections of the document.

1.1 Motivation

Autism is a neurological disorder that affects the normal development of a child. Symptoms occur within the first three years of life and include three main areas of disturbance : social, behavioural and communication, hindering their integration into society and their relationships with others [1]. Nowadays, more people than ever before are being diagnosed with ASD. It is not clear how much of this increase is empty set due to a broader definition of ASD and better efforts in diagnosis. However, a true increase in the number of people with ASD cannot be ruled out. It is thought that the increase in ASD diagnosis is likely due to a combination of the previous factors [2]. In 2010 there were an estimated 52 million cases of ASDs around the world, equating to a population prevalence of one in 132 persons [3]. Figure 1.1 shows the estimated prevalence for ASDs by age group, and gender for some world regions. According to these results, autism disorders are four to five times more common in males when compared with females. The age trajectory follows a similar pattern for males and females, with a sharp rise in prevalence prior to 5 years of age before peaking between 5 and 20 years of age [3].

The most recent worldwide estimations, made in 2012, point to a proportion of 17 in 10000 children with autism and 62 in 10000 with other pervasive developmental disorders in the autism spectrum [4]. In spite of the fact that there are no recent statistics for Portugal, there is a study performed in 2005 that estimates that the prevalence of children diagnosed with ASD, between 7 and 9 years old, is approximately 9 in 1000 children for Continental Portugal and 16 in 1000 for Azores, according to Diagnostic and Statistical Manual of Mental Disorders (DSM)-IV's definition [5].

Even though there is no known cure for ASD, there are many therapies that may help children overcome some difficulties and improve their skills [1]. Some of these therapies include Applied Behaviour

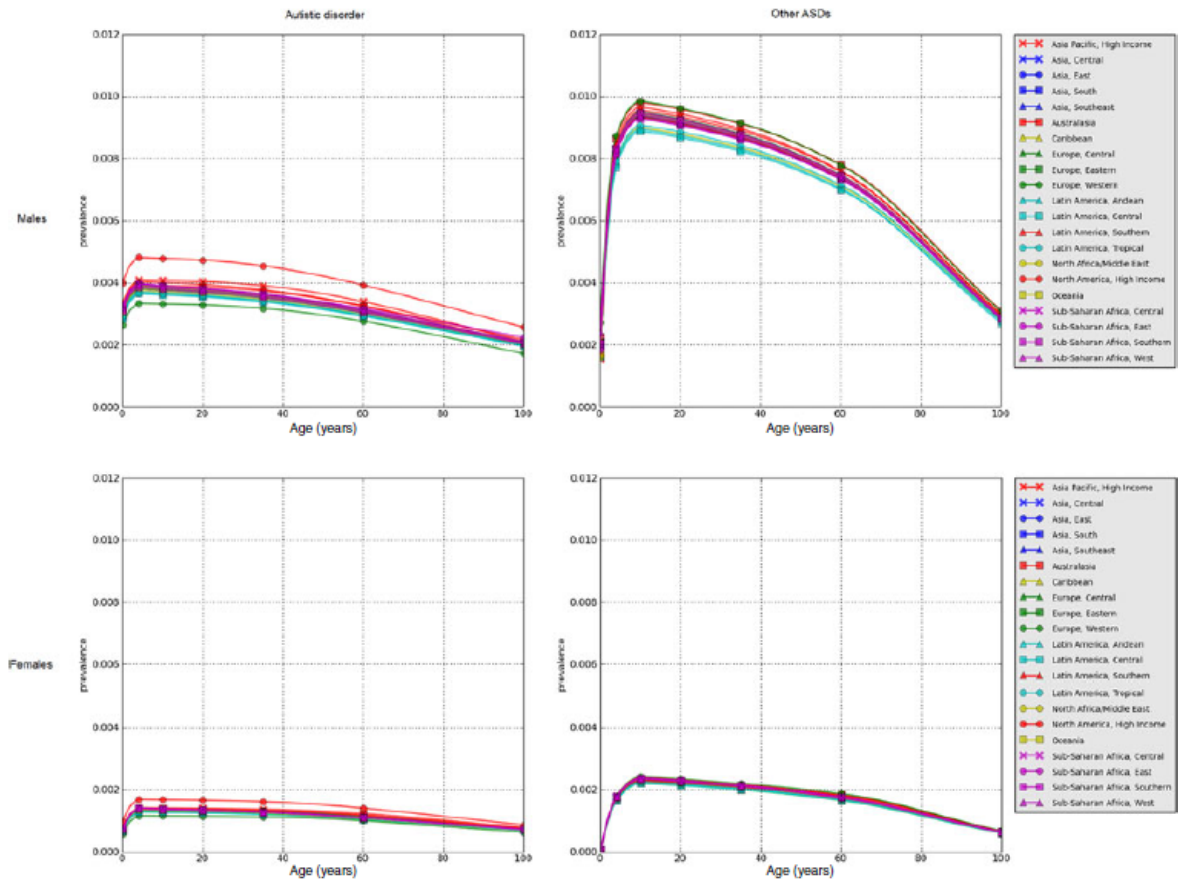


Figure 1.1: Estimated point prevalence of ASDs for male and female in 2010 [3].

Analysis (ABA), Floortime, Son-Rise, Relationship Development Intervention (RDI), among others. It is extremely important having in mind that all children are different so, what is a good solution for one child may not be so good for another. One of the most used therapies is ABA, which relies on the principles that explain how learning takes place, such as positive reinforcement. When a behaviour is followed by some sort of reward, the behaviour is more likely to be repeated [6].

Since autism has no cure, it is extremely important to find an appropriate therapy and treatment program, that may potentially improve the outlook for most young children with autism [7]. Evidence is growing that technology is engaging to many children across the autism spectrum and have been shown to elicit behaviours that may not be seen in child-person interactions [8–10]. Promising methods have shown that the child's natural interests in technology can encourage communication with the therapist [11]. One of the challenges, however, is to help ASD children generalize interaction skills from robots or mobile applications to humans, a connection that is very difficult to establish.

Concluding, in our work, we intend to do an intensive research in the state of the art and making the use of knowledge from several areas of electronic and computer engineering, such as development of a mobile application, artificial intelligence or language processing, in order to improve the social and prosodic skills of a children with autism.

1.2 Framework

The present work was developed in the framework of two projects: Realising and Applied Gaming Ecosystem (RAGE) and Intelligent Networked Robot Systems for Symbiotic Interaction with Children with Impaired Development (INSIDE). This section describes each of these projects and how our work fits in them.

1.2.1 RAGE

The overall objective of this EU project is to develop, transform and enrich the gaming industry, for serious purpose, through the creation of a self-sustainable ecosystem, which is a social space that connects research, gaming industries, intermediaries, education providers, policy makers and end-users¹. This ecosystem provides:

- Business models to support the challenges of the applied gaming industry, promoting new opportunities and value chains;
- A set of practices of the usage of technological components in several real world contexts;
- Centralized access to all resources and services relating to technological components; online training opportunities for developers and educators; an online space to facilitate the collaboration as well as the innovation process between several stakeholders;
- A cloud of advanced technological components, open-source, that allow an efficient development, more appropriate to the desired characteristics of the applied games, which includes: user analysis, emotion detection, user adaptation, generation of credible social behaviour for characters, speech and text processing and gamification.

The RAGE project is a very nice framework for the development of gamified prosodic exercises, such as the ones we intend to develop, as a complement to therapy.

1.2.2 INSIDE

The INSIDE project is funded by Fundação para a Ciência e Tecnologia (FCT) in the framework of the Carnegie Mellon Portugal partnership program. The INSIDE initiative explores symbiotic interactions between humans and robots in joint cooperative activities, and addresses the following key research problems²:

- How can robots plan their course of action to coordinate with and accommodate for the co-actions of their human team mates?
- How can task, context and environment information collected from a set of networked sensors (such as cameras and biometric sensors) be exploited to create more natural and engaging inter-

¹<http://rageproject.eu/>

actions between humans and robots involved in a joint cooperative activity in a physical environment?

INSIDE intends to develop new hardware and software solutions that will support a real-world interaction with children with ASD in a joint cooperative task with therapeutical purposes. Since INSIDE is a project specially dedicated to children with impaired development, and more specifically children with autism spectrum disorders, the type of exercises that we intend to create for these children to develop their prosodic abilities was in fact a request from the therapists.

1.3 Goals

Initially, the goal of this project was the construction of spoken dialogue between autistic children and robots in a hospital environment, in order to develop collaborative learning. However, after an intensive research and through in-loco observation of two pilot tests, we were able to verify that the targeted autistic children had major limitations in dialogue development and sentence construction. This fact justifies the scarcity of speech data from ASD children. This lack of data makes it impossible to carry out any tests and, consequently, the concretization of such a proposal.

In order to come up with an alternative proposal, we meet with therapists who spoke about the need for prosodic training and development, as a complement of therapy, in a home environment. After some research in this topic, we were able to make a new proposal, described below.

This work was developed with the purpose of giving Portuguese children, identified with ASD, a set of exercises that will help them develop and consolidate prosodic skills, both linguistic and non-linguistic. Since their main difficulty is linguistic prosody, it will be the main focus of our work. More specifically, the objectives of the present project are:

- Formulation and implementation of a set of prosodic exercises, with the aim of extending the Virtual Therapist for Aphasia Treatment (VITHEA) - Kids, as a complement of therapy;
- Development of an intonational assessment method, through an imitation task - which is the most challenging aspect of this work;
- Implementation of a mobile application, in android environment, for demonstration and test of the previously referred exercises.

1.4 Document Outline

This document is divided into seven chapters. In Chapter 2 we provide some background knowledge about the concept and history of autism, the main language impairments that characterize this disorder, the diagnosis of ASD in children, and the most commonly used therapies and interventions. Chapter 3 covers related work in terms of technology, namely software for children with ASD, as well as a brief

²<http://www.project-inside.pt/>

survey of intonation validation. Besides, in chapter 4 we explain and show the results for our intonation assessment method. After all investigation was done, in Chapter 5 we describe the development and the implementation of the new exercises that will integrate the virtual therapist. Finally, in Chapter 6, we summarize the highlights of our work and present some future work possibilities.

Chapter 2

Background

In this chapter several topics related to ASD will be discussed, like its definition and history (section 2.1), emotions and speech of ASD children (section 2.2), its tests and diagnosis will be discussed (sections 2.3), and some therapies and interventions (section 2.4) will be described.

2.1 Autism Spectrum Disorder

The "Autism" term was used for the first time at the beginning of the twentieth century, by Eugene Bleuler, to designate a category of thought disorder that was present in schizophrenic people [5]. Three decades later, Leo Kanner, a child psychiatrist, studied the behaviour of a group of children who had in common specific clinical characteristics, never documented before. Thus, in 1943, he published a new clinical syndrome named "Autistic disturbances of affective contact", that described with huge detail 11 children with autism (8 boys and 3 girls, with ages between 2 and 8 years old)[12]. These children had in common major deficits in a daily basis social interactions, whether with parents and family, either with their mate, choosing to be alone. Their behavior was bizarre, characterized by restricted, repetitive and strange interests and activities. The language was peculiar, three children did not talk at all, and the others did pronouns exchanges or literal interpretation of the verbal information, being extremely difficult to hold a conversation. Often, they had intense and disproportionate fears to everyday noises, such as the Hoover or food mixer noise [12]. They seemed to have a great memory, since they easily memorize poems and music and had special interests in numbers or letters. This syndrome, that combined autism, obsessions, speech problems and lack of interaction with people, was, for the first time, distinguished from schizophrenia. A year later, the pediatrician Hans Asperger, described a group of children with the same type of disturbances, such as difficulties in social interactions, restrictive range of interests and repetitive behaviors. However, his observations differed from those made by Kanner in the sense that the children he described displayed a typical development in what concerns to cognitive and language skills [13]. Since 1980, autism became part of the Diagnostic and Statistical Manual of Mental Disorders, in its third edition (DSM-III). This edition included autism in a new class of global pervasive developmental disorders, designated as Pervasive Developmental Disorders (PDD). This category cov-

ered a set of clinical disorders, with early beginning, affecting at the same time a huge range of basic fields of behaviour and development [5, 14]. The Asperger's Syndrome was only featured in the fourth edition (DSM-IV) [15]. In accordance with this edition, people displaying autistic behaviour would receive the separate diagnosis of Autism, Asperger's Pervasive Developmental Disorders - Not Otherwise Specified (PDD-NOS) [14, 15]. Concerning the most recent edition (DSM-V), released in 2013, all autism disorders were converted into one spectrum diagnosis of ASD, no longer being divided into different subtypes. In DSM-V, is defined the following criteria for an ASD to be diagnosed [1]:

- A. " Persistent deficits in social communication and social interaction across multiple contexts ". This can be manifest by the following:
 - (1) Deficits in social-emotional reciprocity, like abnormal social approach and failure a normal conversation through reduced sharing of interests and emotions, that may result in a total lack of initiation of social interaction;
 - (2) Deficits in non-verbal communicative behaviours used for social interaction, ranging from poorly integrated-verbal and non-verbal communication, through abnormalities in eye contact and body-language, or deficits in understanding and use of non-verbal communication, to total lack of facial expression or gestures;
 - (3) Deficits in developing and maintaining relationships, appropriate to developmental level that can vary between difficulties adjusting behaviour to suit different social contexts through difficulties in sharing imaginative play and in making friends to an apparent absence of interest in people.
- B. " Restricted, repetitive patterns of behaviour, interests, or activities ", manifested at least by two of the following:
 - (1) Stereotyped or repetitive speech, motor movements, or use of objects;
 - (2) Excessive fixation on routines, ritualized patterns of verbal or non-verbal behaviour, or excessive resistance to change;
 - (3) Highly restricted, fixated interests that are abnormal in intensity or focus;
 - (4) Hiper- or hipo-reactivity to sensory input or unusual interest in sensory aspects of environment.
- C. " Symptoms must be present in the early developmental period "
- D. " Symptoms cause clinically significant impairment in social, occupational or other important areas of current functioning "
- E. " These disturbances are not better explained by intellectual disability (intellectual developmental disorder) or global developmental delay ". Intellectual disability and ASD often co-occur, but a co-morbid diagnosis of ASD and intellectual disability requires that social communication is below than expected for general developmental level.

2.2 Emotions and Speech

Despite the universality of language impairments in autistic children, the disorder is not characterized by a unitary language deficit. Phonological, lexical, semantic, and syntactic deficits vary widely in children with ASD, with some exhibiting close to normal abilities while others show profound impairments [16].

Language skills in school-age children with autism are excellent predictors of current function and an important predictor of future outcome. Early precursors to speech and language have been well documented in typically developing infants have [17]:

- The ability to discriminate among the phonetic units of speech;
- A keen interest in spoken language;
- The ability to learn from exposure to language.

Children with ASD may have difficulty developing language skills and understanding what others say to them. They also may have difficulty communicating nonverbally, such as through hand gestures, eye contact, and facial expressions [17]. Below we describe some patterns of language use and behaviours that are commonly found in children with ASD.

- **Repetitive or rigid language:** Often, children with ASD who can speak will say things that have no meaning or that seem out of context in conversations with others. Some children with ASD speak in a high-pitched or singsong voice or use robot-like speech. Other children may use stock phrases to start a conversation. For example, a child may say “My name is Tom,” even when he talks with friends or family.
- **Narrow interests and exceptional abilities:** Some children may be able to deliver an in-depth monologue about a topic that holds their interest, even though they may not be able to carry on a two-way conversation about the same topic. Others have musical talents or an advanced ability to count and do math calculations.
- **Uneven language development:** Many children with ASD develop some speech and language skills, but not to a normal level of ability, and their progress is usually uneven.
- **Poor non-verbal conversation skills:** Children with ASD often are unable to use gestures, such as pointing to an object, to give meaning to their speech. They often avoid eye contact, which can make them seem rude, uninterested, or inattentive.

Impairments in social interaction in ASD are frequently observed as a limited use of expressions, and a lack of social and emotional reciprocity. Research has documented that children with ASD are less capable of coordinating social cues, perceiving others’ moods, and anticipating other’s responses [18]. Understanding emotions is a key element in social interactions, since it enables individuals to accurately recognize intentions of others and fosters appropriate responses. Because of the core deficits in ASD involve impairments in reciprocal social interactions and social behaviours, several studies have

investigated emotion recognition. The findings suggest that emotional competence in ASD may be dependent on age, intelligence, and context [19]. Besides this statement, children with autism failed to look at adults, expressing distress, fear, and discomfort and also appeared less concerned when an adult express distress than typical developed children [19].

2.3 Tests and Diagnosis

Nowadays there is not a specific medical test that can diagnose autism, nor any that can distinguish all the sub-groups within ASD. However there are some signs for which parents or caregivers should be aware right from the start of a children's life. These signals, listed below, are designated as “ red-flags ”, and can indicate that a child may be at risk of an ASD [20].

- No big smiles or other warm, joyful expressions by six months or thereafter;
- No back-and-forth sharing of sounds, smiles or other facial expressions by nine months;
- No babbling by 12 months;
- Appears to tune-out or switch-off from other people;
- No back-and-forth gestures such as pointing, showing, reaching or waving by 12 months;
- No words by 16 months;
- No meaningful, two-word phrases (not including imitating or repeating) by 24 months;
- Any loss of speech, babbling or social skills at any age;
- Has a number of unusual preoccupations and attachments;
- Has very short attention span.

This set of signs culminates in the characterization of autism, which is experiencing difficulties with social interactions, lack of communication and some tendency to repetitive behaviours. The image 2.1 demonstrates, in a very clear way, how symptoms and their severity vary widely across the three core areas.

After the symptoms are verified there are some steps to be taken, as can be seen below:

- A. Parents have concerns about their children or a close friend; a professional indicates possible concerns;
- B. Parents should contact their family doctor to discuss their concerns;
- C. If the doctor has the same concerns about the child's development, he will provide a referral to a developmental paediatrician;

- D. The paediatrician will complete a child and family history, examine the child and discuss the concerns with the parents. If there are no doubts, the doctor may give the diagnosis of autism or ASD. Then, the doctor will refer the family to a psychologist for a more detailed diagnosis and assessment.
- E. Parents attend a multidisciplinary team assessment with their child in-depth testing and assessment. This assessment will provide detailed information about the child's autistic characteristics, degree of disability and perceived functional abilities. Then the family will be assigned a case manager from the assessment team who will provide ongoing support and assistance.

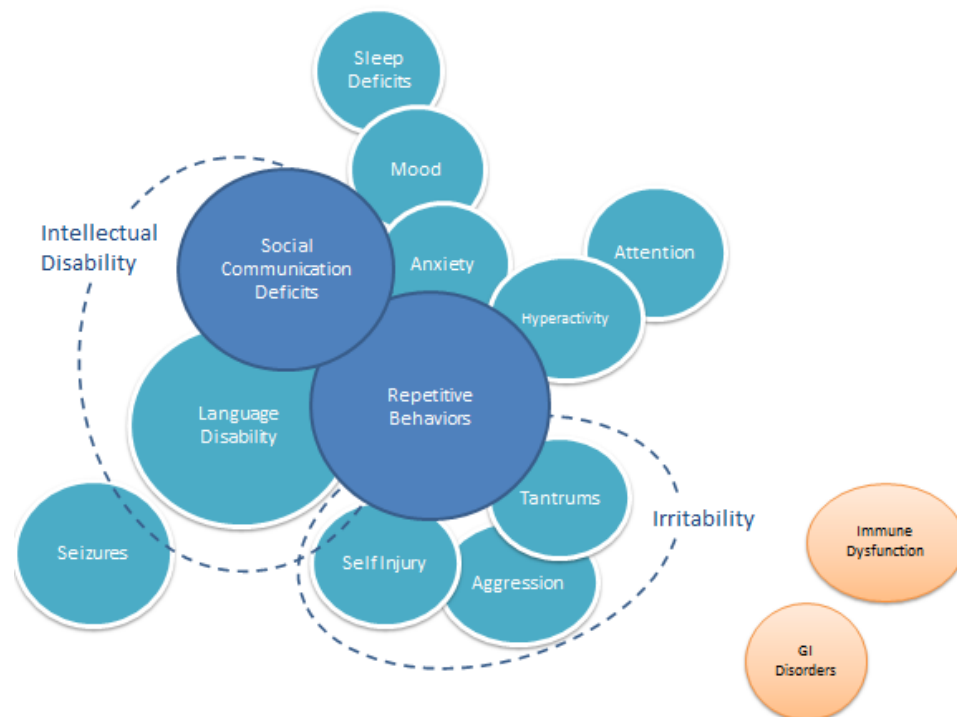


Figure 2.1: Autism Symptoms ¹.

2.4 Therapies and Interventions

ASDs are lifelong chronic disabilities. At this moment, there is no cure for the core symptoms of autism. However, there are several therapies that can help an individual to have a better quality of life and are scientifically proven to improve communication, learning and social skills. Below we describe some therapies that are commonly applied to individuals diagnosed with ASD.

2.4.1 Applied Behaviour Analysis (ABA)

Nowadays, most of the therapies used for ASD are based on ABA. Since the early 1960's, ABA has been used by hundreds of therapists to teach communication, play, social, academic, self-care, work

¹<http://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH/Autism.html> (last visited on 12/12/2015)

and community living skills, and to reduce problem behaviours in learners with autism [21]. There is a great deal of research literature that has demonstrated that ABA is effective for improving children's outcomes, especially their cognitive and language abilities. ABA methods use the following three step process to teach [21]:

- An **antecedent**, which is a verbal or physical stimulus such as a command or request. This may come from the environment or from another person, or be internal to the subject;
- A resulting **behaviour**, which is the subject's (or in this case, the child's) response or lack of response to the antecedent;
- A **consequence**, which depends on the behaviour. The consequence can be classified as positive or negative whether they add or remove a certain stimulus in order to improve a certain behaviour. Consequence based approaches can also be grouped in punishment based approaches (if they consist of adding or removing an aversive stimulus) or reinforcement based approaches (if the manipulated stimulus is something that pleases the individual).

There are innumerable therapies that use the assumptions of the ABA method, but with different emphasis and techniques. Some of these therapies are:

- **Pivotal Response Treatment (PRT)** – This technique is used to teach language, decrease disruptive behaviour and increase social, communication and academic skills by focusing on critical (or pivotal) behaviours that, consequently, affect a wide range of behaviours. The critical behaviours are considered to be motivation and initialization of communication with others. This type of therapy is child-directed and the goal is to produce positive changes in the pivotal behaviours and to give the child the capacity to monitor his or her own behaviour [6, 22].
- **Verbal Behaviour (VB)** – Therapy that uses the analysis of the method of ABA, Skinner, as a basis for teaching language and shaping behaviour. This technique is designed to motivate a child to learn language by developing a connection between a word and its value. The objective of this kind of interventions is to develop functional language skills, in particular the following core functional units identified by Skinner [23]:
 - **Mand:** A way of response that is not controlled by an antecedent stimulus. For example, the child learn how to say the word “ cookie ” when they are really interested in obtain a cookie.
 - **Tact:** Response that is determined by an antecedent stimulus. Using the previous example, in this case, the child say the word “ cookie ” after seeing a cookie.
 - **Echoic:** A response which is determined for a certain action in a different person. For example, when a child say the word “ cookie ” after seeing someone eating a cookie.
 - **Intra-verbal:** Similar response to the echoic one, but without a point-to-point correspondence. An example of this kind of response is when someone say the word “ cookie ” in response to a question, for example, “ What do you want? ”.

- **Early Start Denver Model (ESDM)** – It is a developmental, relationship-based intervention approach that utilizes teaching techniques consistent with ABA. The goals are to promote social skills – communicative, cognitive, and language – in young children with autism, and reduce abnormal behaviours associated with ASD. This technique is mostly used for babies in the first months of development, up to twelve months, with symptoms of ASD.

2.4.2 Floortime

The premise of Floortime, suggested by Greenspan, is that an adult, in particular the parents, can help a child expand his circles of communication by getting down on the floor with the child to engage him at his level [24]. This is the reason why this therapy is called Floortime. It is a type of intervention that is focused on child's strengths and interests and the goal is to help the child reach six developmental milestones that contribute to emotional and intellectual growth [24]:

- Self-regulation and interest in the world
- Intimacy or a special love for the world of human relations
- Two-way communication
- Complex communication
- Emotional ideas
- Emotional thinking

2.4.3 Son-Rise

This program was created by the Kaufmas, also creators of The Option Institute and Autism Treatment Center of America. This program treats individuals diagnosed with ASD by embracing the physical and emotional expressions of autism with care and understanding. The main objective is to enter in the individual world, through enthusiastic and playful interactions and then, guide them into our world, like floortime ².

2.4.4 Relationship Development Intervention (RDI)

RDI is a system of behaviour modification through positive reinforcement, as a parent-based treatment using dynamic intelligence, developed by Dr. Gutstein. The main objective of this therapy is to improve the long-term quality of ASD individuals, by helping them improve their social skills, adaptability and self-awareness [25]. Its six objectives are [25]:

- **Emotional Referencing:** The ability to use an emotional feedback system to learn from the subjective experiences of others.

²<http://www.stanleygreenspan.com/> (last visited on 22/12/2015)

- **Social Coordination:** The ability to observe and continually regulate one's behaviour in order to participate in spontaneous relationships involving collaboration and exchange of emotions.
- **Declarative Language:** The ability to use language and non-verbal communication to express curiosity, invite others to interact, share perceptions and feelings and coordinate your actions with others.
- **Flexible Thinking:** The ability to rapidly adapt, change strategies and alter plans based upon changing circumstances.
- **Relational Information Processing:** Relational Information Processing: The ability of solving problems that have no “right-and wrong” solutions.
- **Foresight and Hindsight:** The ability to reflect on past experiences and anticipate potential future scenarios in a productive manner.

2.4.5 TEACCH

TEACCH is an evidence-based service, training and research program for individuals with ASD. This approach includes a focus on the person with autism and the development of a program around the person's skills, interests, and needs. The priorities of TEACCH are center on understanding autism, making necessary adaptations, and selecting strategies for intervention that utilize the person's existing skills and interests, as written above [26]. The educational program recommended by this system is founded on the following principles:

- Strengths and interests;
- Ongoing assessment;
- Assistance in understanding;
- Parent collaboration;
- Individualization.

2.4.6 SCERTS

This approach is an educational models that uses some principals of most of the approaches previously described. SCERTS is more focused with helping children with autism to achieve “Authentic Progress,” which is defined as the ability to learn and spontaneously apply functional and relevant skills in a variety of settings and with a variety of partners. As its own noun indicate, this is focused on [27]:

- “SC” Social Communication - Development of spontaneous, functional communication, emotional expression and secure and trusting relationships with children and adults.
- “ER” Emotional Regulation - Development of the ability to maintain a well-regulated emotional state to cope with everyday stress, and to be most available for learning and interacting.

- “ TS ” Transactional Support - Development and implementation of supports to help partners respond to the child’s needs and interests, modify and adapt the environment, and provide tools to enhance learning.

2.4.7 PEPS-C

Despite not being a therapy, it is an intervention that we could not forget to talk, since it is the most used tool for evaluation of prosodic skills of children diagnosed with ASD, the Profiling Elements of Prosody in Speech - Communication (PEPS-C) [28]. PEPS-C is a test that assesses both receptive and expressive prosodic abilities. This procedure has two levels: the form level assesses auditory discrimination and the voice skills required to perform the tasks; the function level evaluates receptive and expressive prosodic skills in four communicative functions: questions versus statements, liking versus disliking, prosodic phrase boundaries, and focus. For a better understating of PEPS-C structure, a scheme was made (figure 2.2).

Levels	Tests	Tasks
	Vocabulary	
Form	Short Items	Discrimination
		Imitation
	Long Items	Discrimination
		Imitation
Function	Interaction	Reception
		Expression
	Affection	Reception
		Expression
	Segmentation	Reception
		Expression
	Focus	Reception
		Expression

Figure 2.2: PEPS-C Structure.

This assessment was ported to European Portuguese (EP), in 2014, by Marisa Filipe [29]. In order to meet the EP characteristics several modifications were made, mainly on the auditory stimulus used.

2.4.8 PECS

Nowadays, the most inspiring method, while developing software for children with ASD is the traditional Picture Exchange Communication System (PECS). PECS is an augmentative communication system, developed to help subjects in quickly acquiring a functional means of communication [30]. It is versatile and inexpensive, since tutors can create and print the pictures that the children needs, as we can see in figure 2.3.

2.5 Summary

In this chapter, we provided some background on ASD and related concepts, as well as therapies and interventions followed by therapist in order to improve autistic people skills, so that the reader could






















 I want		 I see		 thank you	
 drink	 biscuit	 apple	 cake	 crisps	 banana
 book	 sand	 bricks	 pens	 farm	 puzzle
 shoe	 jumper	 trousers	 coat	 sock	 hat

Figure 2.3: Example of PECS.

be familiarized with such disorder.

Accordingly to DSM-V, an individual can be diagnosed with ASD if his/her behaviour shows persistent deficits in social communication and interaction, restricted and repetitive behavior and interests, and these symptoms occur in the early developmental period.

In order to minimize ASD impairments and difficulties, there are several approaches, however the most common one nowadays is ABA, which is a therapy characterized by its focus on targeting a certain behaviour, as well as events that precede and follow such behaviour. However, there are also approaches that focus on the individuals' strengths and preferences, like Floortime and the Son-rise therapies.

We finalized the present chapter by referring PEPS-C, which briefly is a test that evaluates prosodic skills of children diagnosed with ASD, and PECS which is the most inspiring approach for software development for autistic children.

Chapter 3

Related Work

As described in the previous chapter, ASD comprises impairments in communication and development of verbal language. Another important statement to have in mind is that some children are non-verbal or only acquire verbal skills later than their typically developing peers [1], so they need alternative means of communication in order to be able to express themselves. Additionally, verbal communication and also prosodic characteristics, could possibly be improved if autistic children were given tools for development of language skills that took their needs into account.

Having this in consideration, there are two main fields of research and development of a wide range of techniques that help therapists, as well as caregivers of autistic children, and that may have a huge impact in the development of their social skills. The main objective in studying these areas, is to help children achieve some milestones. These fields are robot-children interaction, and the development of mobile applications.

This chapter is structured in two main topics that are related to our goals: the use of technology for children with ASD, and surveys related to intonation assessment.

3.1 Technology for children with ASD

In this section we describe and analyse works that, as described before, study the use of technology to teach children with ASD: characterization of the user needs, software for autistic children and virtual therapists. Afterwards, we will discuss the findings of such works and how they can be useful to our work.

3.1.1 Characterization of user needs

In order to develop software that better meets the needs of users with cognitive impairments such as ASD, several authors performed surveys regarding aspects such as software features preferred by children with ASD and their caregivers, namely: previous experiences, main difficulties that should be addressed, and reasons to abandon previously adopted technology. For this specific project, it is important to analyse desirable software features, so that our application may be acceptable for this specific

audience.

As it is known, people with autism form a diverse group, however, there are commonalities which may be taken into account when designing software. In [31] it is advised that a learning environment for children with autism, should be as dependable and predictable as possible, with any required unpredictability carefully introduced in a controlled way. They also advise that the learning activities should be challenging, but children should not be penalized for mistakes. Lastly, it is recommended that children should be allowed time to enjoy their mastery of a skill before moving on to the next activity. Analysing several surveys we came across some design issues that we should take into consideration while developing a software for ASD children:

- Children with autism generally are highly visual, so it is important to present narratives and photo-narratives as pictures, with no verbal commands [32].
- They tend to employ local rather than global integration; a child with autism might focus exclusively on some seemingly irrelevant detail [33], so the screen design might be very simple.
- They may be highly sensitive to noise, finding intolerable noise which is barely perceptible or unremarkable to others [34]; hence sound features should be moderately used.
- They may find failure very debilitating [31]; they should not be penalized for a wrong answer, indeed they should be encourage to try again.
- They generally enjoy repetition [32].

In a study through questionnaires, several suggestions were made by the participants, such us: taking into account sensory integration issues; making products portable and easier to use (e.g., using voice activation); incorporating fun elements [35]. Another similar survey suggested three fundamental features: device portability, ease of use (and ability to evolve along with the user) and ease to upgrade or replace [36].

3.1.2 Software for children with ASD

With technology becoming more important and useful every day, a variety of software and studies aimed at children with autism is now available. The range of such software can go from entertainment and planning to the development of communication, academic, social or emotional skills ¹. In this section the focus will be the development of the vocabulary skills (works that try to increase the number of words that children with ASD recognize and use) and the communication in a social context (authors that try to deliver some examples to children on how to socialize and express themselves with emotions, allowing them to learn how to react and feel more comfortable when confronted in real life). At the end, we will describe some commercial tools, available for purchase or free download, for children with autism or other type of impairments, with no research associated, and having only the purpose to answers the users' needs.

¹<https://www.autismspeaks.org/autism-apps>

Development of Vocabulary Skills

Several studies are focused on developing and increasing children's vocabulary, by using a specific type of stimuli that motivate children.

Moore and Calvert (2000) [37] conducted a study focused on stimulating children with visual and auditory stimuli, which analysed the impact of computers on the extension of the vocabulary of the child, by developing a software program that builds upon behavioural learning principles to enhance the vocabulary skills of children with autism. The software provides animations and interesting sounds as a reinforcement when the child provides correct responses to the given commands. The results are positive, showing that children pay more attention when using the computer than with the tutor, and that they felt more motivated to continue the study using the computer. Besides, an important conclusion is that the number of correct words identified increased.

Using a different approach, **Bosseler and Massaro** realized two different studies, one in **2003** [38] and another one in **2006** [39]. For both studies they used an animated character called Baldi as 3D language tutor, to help in learning speech and reading through the association between pictures and spoken words. In the first study, they asked children to take pictures of objects and surroundings at home, which were then incorporated in the lessons. The results showed that the number of vocabulary words increased, and it was concluded by the authors that the positive results were due to the lessons with Baldi. On one hand, using an embodied agent might be a positive reinforcement since can more easily relate with a non-human character. On the other hand, with time children may get bored with the agent itself, leading to loss of interest. This could mean that over time the lessons may be increasingly strained, which ultimately leads to decreased motivation and possibly to the reduction of visible results.

For the second study the program provided exercises in every lesson with a unique set of items appropriate for the children's vocabulary, knowledge, and abilities. In each lesson they could use the Baldi agent or only the spoken word. During the exercises the child had to choose the correct image given the word that was said, or to choose certain areas of the image according to what was asked. In some exercises they had positive reinforcements with a happy face, or a sad face for incorrect answers. At some point, the child was asked to vocalize the name of the item or its function. Results showed that there were more correct answers with Baldi, than with the audio alone and overall the correct answers increased.

Both studies show that it is possible to test several approaches with the same tool. Despite the fact that the problems listed above are a reality in these studies, the results are clear in stating that using the agent is more beneficial than without it.

Hetzroni and Shalem(2005) [40] tried a simpler approach focused on vocabulary development, after a study targeted at the development of communication skills in a social context. The authors developed a tool to implement a 7-step gradual fading protocol, where at first it shows the picture of a food item with the name on it, then the pictures fade out until only the name is visible. Emotions were used as

reinforcement. Later, in the classroom setting, children were asked to point out the food items that were written. Results show that correct matches between the text and the food items improved for all participants, but the software shows limited resources.

Finally, **VITHEA (2013)** [41]² is an international awarded platform, designed for aphasia patients and therapists, which comprises two modules. The patient module contains a set of exercises, in which the patient needs to orally reply to a certain stimulus; the answer is then recorded and matched to a set of one or more correct answers. The administration module, allows therapists to create and manage exercises, users and the multimedia resources to be used in the exercises, as well as to check each patient's information and exercise statistics. Despite the fact that **VITHEA-KIDS (2015)** [42] is based on the same infrastructure of VITHEA, it makes use of a different type of exercises: multiple choice exercises, which target vocabulary acquisition and/or improvement of generalization skills, and the targeted users are children with ASD. These exercises are composed by a question, a stimulus, that could be a picture or text, and a set of possible answers (textual or pictures, respectively), in which only one of the answers is correct. Besides, this platform allows caregivers to build customized multiple choice exercises while taking into account specific needs/characteristics of each child. In figure 3.1, an example of an exercise of the child's module of VITHEA - KIDS is represented.



Figure 3.1: VITHEA-KIDS platform [42].

Despite not being directly related with the development of vocabulary skills, it is important to refer a study made by **Thorson et al. (2016)** [43], since it is a study focused on the development of prosodic abilities. This survey consists of a procedure (AP: Assessment of Prosody) for assessing basic prosodic perception and production abilities of minimally to non-verbal children and adolescents with ASD. The procedure consists of three primary sections. The first is an Optional Primer Phase, the second is the Learning Phase, and the third is the Assessment Phase. The assessment methodology is basically introducing the concept of low and high, first via animal sounds and then via human speech. The high location is represented by a bird with corresponding bird chirp sounds. The low location is represented by a cow with a corresponding cow moo sound. Cartoon birds are present in each of the top two squares on the magnetic board, and cartoon cows are present in the two bottom squares. The items are first introduced and then the animal sound low/high association task begins with a series trials. For human speech the procedure is the same, but without the animal picture for stimuli. This procedure is extremely important since it tests the prosodic skills that are necessary for music-motor based intervention thera-

²<https://vithea.l2f.inesc-id.pt>

pies and provides a baseline for comparing how these types of therapies impact prosodic components of the language system.

Communication in a Social Context

Parsons, Leonard and Mitchell conducted several studies on the use of virtual environments for social skills training. In [44] they investigated whether two subjects with ASD could interpret virtual scenes meaningfully and if they could provide an appropriate social response. The results show that each participant was able to remember and properly execute social knowledge gained during the study. In [45] they used a virtual coffee to teach the learning objective "finding a place to sit" in different circumstances. The students showed improvements in social judgements and awareness of social reasons of why they might do something in a certain context. In [44] twelve individuals with ASD were matched with two other pupils, one on verbal IQ and the other on performance IQ. The ASD group performed on par with their matched counterparts, and the majority of the group seemed to have a basic understanding of the virtual environment as a representation of the real world. More recently, this study was expanded to include video clips of real world exemplars as measures for generalization, [46]. Participants made improvements in judgement and reasoning about where to sit both in the video examples. They found that among the participants who showed the greatest improvement in social reasoning were two participants with the lowest verbal IQ scores. Furthermore, the participants were not encouraged or instructed to transfer what they had learned in the virtual coffee to the video clips, therefore suggesting that spontaneous generalization may have the potential to occur.

De Leo and Leroy (2008) [30] designed and developed a software based on PECS that offers a new media for communication and socializing while overcoming the shortcomings with paper pictures. The tool developed by them is called PixTalk³ and allow the child to choose and combine images to form a message. Although we were not able to have access to the results of this study, we thought that was important to make a reference since this tool was developed for children with communication impairments.

Cihak et al. (2012) [47] tried a simple approach consisting on the evaluation of PECS used in conjunction with Video Modelling (VM), as a technique to increase independent communicative interactions in children with limited to no verbal communication skills. In this study, the authors used video to teach what would normally be taught with PECS. Tests show that as a result of using VM in conjunction with PECS, all students increased independent communication interactions, and the student's rate of learning was quicker when using VM. Still, this technique is very limited in giving the children a comprehensive set of tools to enable learning of communication skills, and communicating with others.

Ohene-Djan (2010) [48] developed Winkball, a new Internet-based video messaging and broadcasting technology, designed to support the teaching of oral and visual communication skills in schools. Despite not targeting children with ASD, the author believes that the technology can also be beneficial

³<http://www.communicationautism.com/>

for this population, since the use of media is usually perceived as interesting and motivating. The author does not show any tests or results related with the tool, and therefore it is difficult to draw conclusions about its effectiveness.

Commercial Tools

In the market there are a few applications that, despite not being the matter of a study for which we can assess experimental results are still important to look and evaluate. In this section our focus will be commercial tools for learning and acquiring feelings and emotions, as well as some communication skills.

- **Learning with Rufus - Feelings and Emotions**⁴: Paid application that uses a child-friendly character to teach emotion words, facial expressions associated with emotions such as happy, sad, anger, and others, and to identify emotions in others (figure 3.2). This application is organized into three parts, a learning phase and two games:

- Learning phase - A preview of the facial expressions is shown to the child before the game starts.
- Find It! - A number of facial expressions are shown and the child is directed to select a specific emotion.
- Name It! - A single facial expression is shown and the child is asked to name the emotion.

The game is highly customizable to meet the needs of children with varying skills, ability levels, and learning styles. Besides, in order to keep children interested and motivated, a variety of features such as reward sets, positive reinforcement, music, sounds and toy break are included.

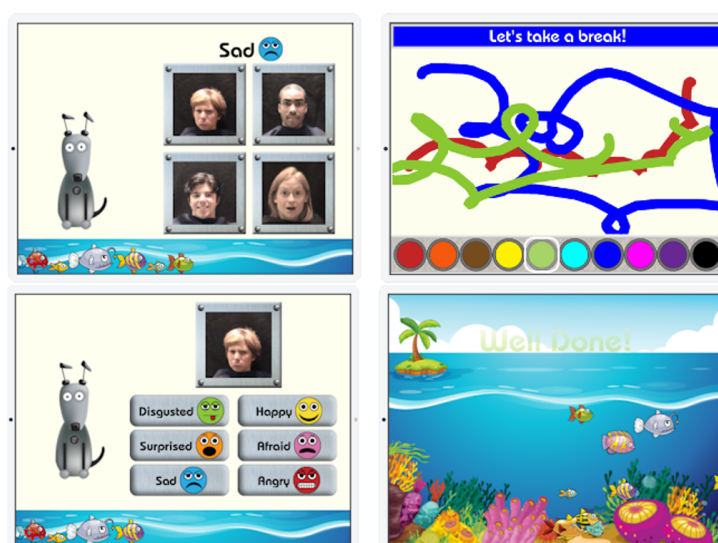


Figure 3.2: Learning with Rufus.

⁴<http://www.rufusrobot.com/emotions.php>

- **Emotions and Feelings - Autism**⁵: This app includes a social story about different emotions and feelings we may have throughout the day, and a simple visual support for asking how someone is feeling, or identifying feelings or emotions. The start menu lets the user choose to either read a story or go to the "Emotions and Feelings" page. The story describes different emotions and feelings and what may cause them and the "Emotions and Feelings" page has nine buttons with different emotions and feelings from the story that say the word when pushed. This is a paid application and it is possible to see its layout in figure (figure 3.3).

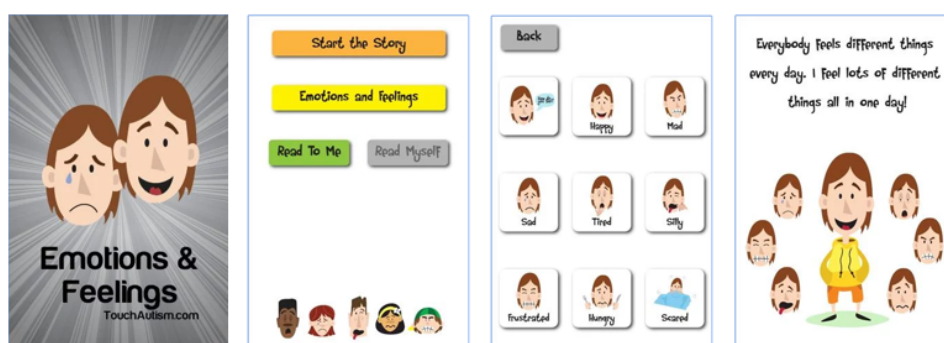


Figure 3.3: Emotions and Feelings - Autism.

- **SPEAKall**⁶: It is a paid APP available for iOS and android. This is an evidence-based app specifically designed to introduce Argumentative and Alternative Communication (AAC) in ASD and/or developmental speech and language disorders. This application has been special designed to help children and adults with little to no functional speech acquire an initial symbol vocabulary and learn how to construct sentences. It can be used to target a variety of communication goals such as functional communication, natural speech production, emerging language and social interaction and social-pragmatic skills. Its interface is customizable to each learner's specific needs. In image 3.4 a screenshot of this app is presented.

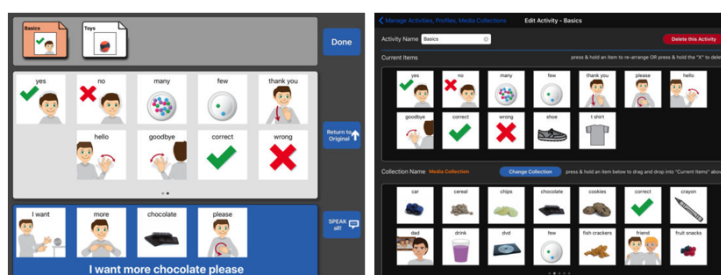


Figure 3.4: Screenshots of SPEAKall application.

- **iCommunicate**⁷: Paid app available for iOS and it lets the user to design visual schedules, storyboards, communication boards, routines, flash cards, choice boards, speech cards, and more. This app uses the PECS format to communicate. In image 3.5 a screenshot of this app is presented.

⁵<https://play.google.com/store/apps/details?id=com.TouchAutism.EmotionsFeelings>

⁶<http://speakmod.com/speakall/>

⁷<http://www.grembe.com/icomunicate>

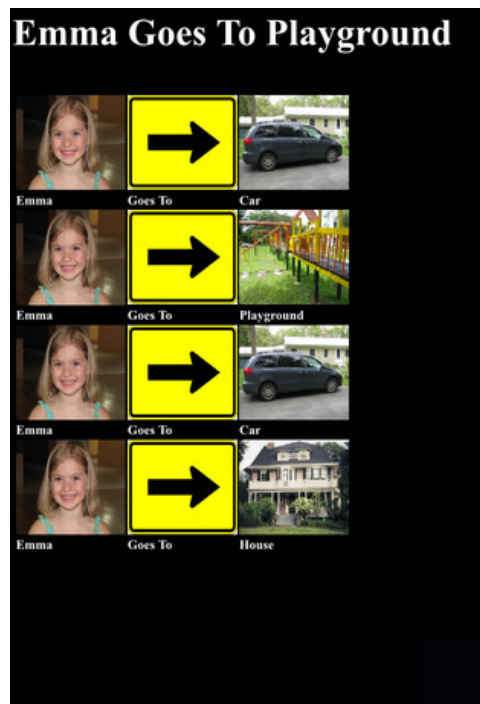


Figure 3.5: iCommunicate application.

- **TalkInPictures**⁸: This free application allows to build informative or requesting sentences using pictures and audio only, as well as to add new sub-menus and terms. The text captions are hidden from the user until the sentence is built and are converted to audio through text-to-speech synthesis, which is available for many languages, including Portuguese. The interface of this app is showed in figure 3.6.

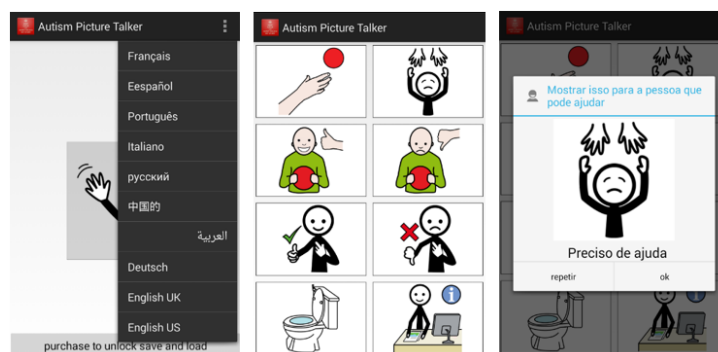


Figure 3.6: TalkInPicture application.

3.1.3 Discussion

Overall, most of the approaches and methodologies proposed so far have, as main objective, the self-development. The majority of the proposed work is inspired in PECS, which is a simple system that children understand, is simple to use and it is even possible to customize, although the possibilities are quite limited in most of the interventions.

⁸<https://play.google.com/store/apps/details?id=com.androidinlondon.autismquicktalk>

Summing up, the main findings provided by the surveys reviewed in subsection 3.1.1, were the fact that device portability is very important, as well as the content customization. Besides, we should always take into account the needs and preferences of children diagnosed with ASD.

Regarding subsection 3.1.2, in table 3.1, we tried to summarize the most common and studied tools as well as their features. In the development of vocabulary skills field, we can conclude that social communication continues to be missing, probably to avoid stressing the children. Besides, we can also conclude that media support is found to be very important; videos still do not seem to have much importance, since most of the studies do not support or use them in tests, however many studies support images, audio and animations. Finally, most of the tools enable the possibility of changing the content, which may suggest that researchers realized the importance of adapting the tool to the user, in order to achieve better results. These results also apply to the studies that focus on developing communication in a social context, being the only obvious difference, the inclusion of message exchange in two of the studies [30, 48]. Other commercial tools include more features supporting most media formats, messages and the possibility to change a few aspects of the content and some features of the tool it self (a feature not found in any of the other studies).

All the studies explore the use of multimedia, which shows us that this is the most effective way of having the attention of the children and, consequently, help in the development of new communication skills. In general, we see that the vocabulary expansion field is using a more mechanic approach, supported by imitation combined with a visual and audio positive reinforcement. This kind of approach, despite helping children to talk, does not help them to communicate with other people and to express themselves with emotion. These tools may also have a high probability of leading to a lack of interest over time, since the contents and interface are always the same in most of the reviewed interventions.

New alternatives have emerged with touch screen technologies (e.g. smartphones and tablets), bringing new opportunities to users(usually paid), and mostly designed to help parents in the interaction with their children. Besides, taking in advance, with the use of such technology, it is possible to explore and develop other capacities such as prosody. Relating with this field, the commercial tools available nowadays help children to understand the differences between intonations and facial expression associated to each emotion, but do not teach the children how to express themselves with emotion, while having a dialogue with someone.

Table 3.1: Summary of Related Work

	Number of Participants	Maintained or Improved Skills	Costume Content	Audio	Images	Video	Animations	Emotional Skills	Reinforcement	Availability
Development of Vocabulary Skills										
Moore and Calvert (2000) [37]	14	✓	✓	✓	✓	x	✓	x	✓	n/a
Bosseler and Massaro (2003) [38]	8	✓	✓	✓	✓	✓	✓	x	✓	n/a
Hetzroni and Shalem (2005) [40]	6	✓	✓	x	✓	x	x	x	✓	n/a
Bosseler and Massaro (2006) [39]	8	✓	x	✓	x	✓	✓	x	✓	n/a
VITHEA-KIDS (2015) [42]	n/a	n/a	✓	✓	✓	x	✓	x	✓	Free
Thorson et al. (2016) [43]	10	✓	x	✓	✓	x	x	x	✓	n/a
Communication in a Social Context										
Pearsons, Leonard and Mitchell (2002) [45]	7	✓	x	✓	✓	x	x	✓	✓	n/a
Pearsons, Leonard and Mitchell (2004) [44]	12 with ASD + 2pupils	✓	x	✓	✓	x	✓	x	✓	n/a
Mitchell, Pearsons and Leonard (2007) [46]	6	✓	x	x	x	✓	x	✓	✓	n/a
De Leo and Leroy (2008) [30]	n/a	n/a	✓	x	✓	x	x	x	n/a	PixTalk
Ohene-Djan (2010) [48]	n/a	n/a	✓	✓	x	✓	x	x	✓	Winkball
Cihak et al. (2012) [47]	4	✓	✓	x	x	✓	x	x	✓	n/a
Commercial Tools										
Learning with Rufus - Feelings and Emotions	n/a	n/a	✓	✓	✓	x	✓	✓	✓	Paid
Emotions and Feelings - Autism	n/a	n/a	x	✓	✓	x	x	✓	n/a	Paid
SPEAKall!	n/a	n/a	✓	✓	✓	x	x	✓	n/a	Paid
iCommunicate	n/a	n/a	✓	✓	✓	x	x	x	✓	Paid
TalkInPictures	n/a	n/a	✓	✓	✓	x	x	x	n/a	Free

3.2 Intonation Assessment

In this section we will describe some works that helped us develop the intonation assessment method. This method is present in a repetition task that will try to improve the intonation skills of autistic children, allowing to express themselves with some emotion, such as express like and dislike or how to express a question or an exclamation. For this purpose, an exhaustive literature search was made and realizing the lack of surveys of intonation validation. So, in order to have surveys about this topic, we had to increase the range of research and study some second language learning systems, that use some methods to validate the pronunciation of a second language learner, being the principal almost the same for intonation validation.

Second language learning systems have two big field of research, being the first pronunciation evaluation and the second nativeness, fluency and intonation evaluation. We will present some surveys for both fields and then we will discuss the findings and how it helped us reach our method for the intonation assessment.

3.2.1 Pronunciation Evaluation

In this subsection we present some surveys related with pronunciation evaluation. **Witt and Young (1997)** [49] developed the Goodness of Pronunciation (GOP). The goal is to have a score of correctness or confidence for each phoneme of a desired transcription for a sentence. The GOP score is finding using two recognition passes of a sentence, the first uses forced alignment to transcriptions determined from a pronunciation dictionary. The second consists of a monophone loop permitting recognition of all possible sentences of phonemes, thus recognising the most likely phoneme sequence. More precisely, this pronunciation score is defined as the ratio of the likelihood of the phoneme which should have been said (forced alignment) and the likelihood of the phoneme that actually has been said (phoneme loop). The results of the evaluation experiments of the described method demonstrate the method's capability of detecting both individual mispronunciations as well as to give a general assessment of which sounds a student tends to badly pronounce.

Franco et al. (1999) [50] proposed an automatic detection of phone-level mispronunciation for language learning. The main objective of such a proposal was to detect specific phone segments that have been mispronounced by a non-native student of a foreign language. For this purpose two different approaches were evaluated. In the first approach, log-posterior probability-based scores were computed for each phone segment, the probabilities were based on acoustic models for each phone segment. For the second approach they used a phonetically labelled non-native speech database to train two different acoustic models for each phone: one model trained with the acceptable, or correct native-like pronunciations and another model trained with the incorrect, strongly non-native pronunciations. For experiments they had a database composed by 130,000 phones in American English uttered in continuous speech sentences, produced by 206 non-native speakers. The results showed that the set of phones that are mispronounced can be detected reliably for both methods. However the second method had better overall results, when compared with the posterior based method.

Franco et al. (2000) [51] created the EduSpeak system, a software kit for development of voice-interactive language education software. For the development they used a native hidden Markov model (HMM) recognizer, that was adapted to non-native speech. In order to do the pronunciation evaluation, they combined scores using a regression tree to generate scores that correlate with scores from human raters:

- Spectral Match: Compare the spectrum of a candidate phone to a native, context-independent phone model.
- Phone Duration: Compare the candidate duration to a model of native duration, normalized rate of speech.
- Speaking rate: phones/sentences.

They highlighted some of its features such as the availability of speaker-independent recognition models for non-native speakers, presenting experimental results that show over 55% relative error reduction for non-native recognition with no degradation for native speakers.

Gupta et al.(2007) [52] patented a method for generating a pronunciation score. The scoring techniques for this particular invention enable students to acquire new language skills, by providing real-time feedback on pronunciation errors. According to the authors, such feedback helps the student focus on key areas where they need improvement, such as phoneme pronunciation, intonation, duration, overall speaking rate and voicing. This invention comprises three main scoring methods:

- Articulation score: computed at the phoneme level and aggregated to produce scores of word level and the complete user phrase;
- Duration score: provides feedback on the relative duration differences between the user and the reference speaker for different sounds or words in a phrase;
- Intonation score: is tutor-dependent and provides feedback at the word level and phrase level. It computes perceptually relevant differences in the intonation of user and the reference speaker. The intonation scoring method operates to compare pitch contours of reference and user speech, and then derive a score. This score reflects stress at syllable level, word level and sentence level, intonation pattern for each utterance, and rhythm. The smoothed pitch contours are then compared according to some perceptually relevant distance measures.

3.2.2 Nativeness, Fluency and Intonation Evaluation

This subsection is focus on surveys that present solutions for nativeness, fluency and intonation evaluation.

Teixeira et al. (2000) [53] presented a method for training a model to assess speakers nativeness similarly to humans without text information. For the survey they constructed a feature-specific decision tree constituted by word stress (duration of longest vowel, duration of lexically stressed vowel and duration of vowel with max f0 (fundamental frequency)), speaking rate approximations, pitch, forced alignment+pitch (duration between max f0 to longest vowel nucleus, location of f0) and unique events (duration of longest pauses, longest words). After the construction of these decision trees, they made a combination (max or expectation) of "posterior probabilities" from decision trees. As for results, it was concluded that pitch-based features do not generate human-like scores (only weak correlation between machine and human scores) and that the inclusion of posterior recognition scores and rate of speech helps considerably (correlation near to .7).

Imoto et al. (2003) [54] introduced a method for sentence-level stress detection of English Assisted Language Learning (CALL) by Japanese students. They also proposed a two-stage recognition level that first detects the presence of stress and then identifies the stress level. Stress models are set up according to stress level, syllable structure and position of the syllable in a phrase. Three streams of acoustic features (pitch, power and MFCC (Mel-frequency Cepstrum Coefficients)) were used, and their weights are estimated by discriminant analysis. Decomposing stress recognition into two stages, one to detect the presence of stress and the other to determine the stress level, represented in the diagram of figure 3.7, improved recognition accuracy together with optimization of the weights at each stage. The results showed that the method achieved an accuracy of 95.1% for native and 84.1% for non-native speakers.

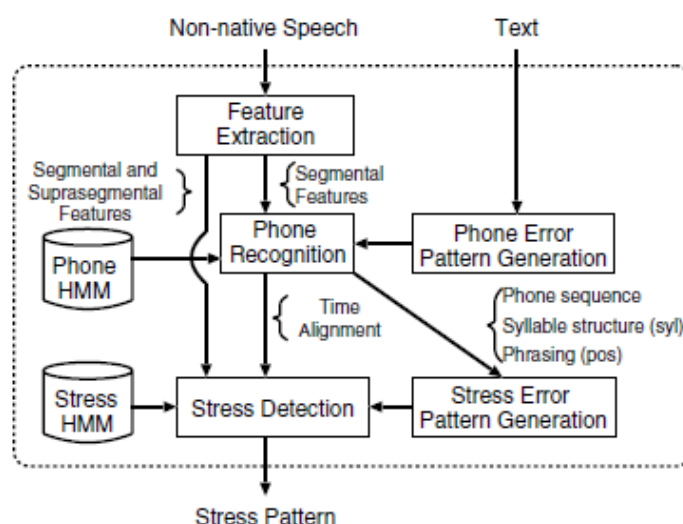


Figure 3.7: Diagram of detection of sentence stress [54].

3.2.3 Discussion

Summing up, the findings provided by the studies in the previous section were the fact that, for pronunciation evaluation, the segmental context and lexical stress of a production determines whether it

is pronounced correctly or not. In order to achieve a pronunciation evaluation there are several features that we might have into consideration, such as phoneme segmentation, speaking rate or pitch contour.

As for nativeness, fluency and intonation evaluation, we conclude that the intonation information can influence the proficiency and understandability of a second language speaker. Besides, proficient second-language speakers often have difficulty producing native-like intonation. For this evaluation the studies that we present had into consideration acoustic features such as pitch, power and MFCC as well as word stress features such as duration of longest vowel, duration of stressed vowel and duration of vowel with max f_0 .

A very recent trend in many speech and language technologies is the use of deep learning approaches. However, this type of approach requires very large training databases, and this is one of the main limitations in our work.

3.3 Summary

In the present chapter, we went through several surveys and works that somehow relate to our goals, essentially in two distinct aspects: the existing software for children with ASD, and the state of the art concerning intonation validation. Summing up, the main contributions provided by our work, are the following:

- Communication and social skills are the most relevant skills wanted to be developed by autistic children, particularly language and emotional skills.
- As for features that would be desirable in technology for autistic children, many suggestions were made, from hardware characteristics, such as device portability, to software features.
- Most of the surveys analysed are only concerned with self-development, and do not give importance to the social aspect of communication.
- Media support, such as images and audio, are very important for children to learn more quickly and be more motivated in pursue all the activities.
- The customizable aspect of tools are very important, since it allows the parents to adapt it to the children tastes.
- The use of touch screen technologies bring the opportunity to explore and develop other skills, helping the children to achieve some communication goals.
- Regarding intonation validation for emotion detection we concluded that there is not studies or approaches concerning with it, so we studied approaches for second language speakers.
- The study of such surveys for second language speakers allow us to understand what are the most important features to have in mind when doing the intonation validation, as well as several approaches that we may consider in our work.

Chapter 4

Intonation Assessment Method

The goal of the intonation assessment method is to evaluate and develop the child skills to imitate different intonations, such as affirmation, question, pleasure and displeasure, for short stimuli (words). For achieving this goal, there was a need to study surveys related to intonation validation but, as previously said, we concluded that there is a gap in the state of the art for this theme. So, in order to overcome this limitation, we studied surveys related with pronunciation training for second language learners, in accordance with chapter 3.2. The architecture of the proposed model is shown in figure 4.1. In the following subsections we will describe each the steps, and then present the final results for this module.

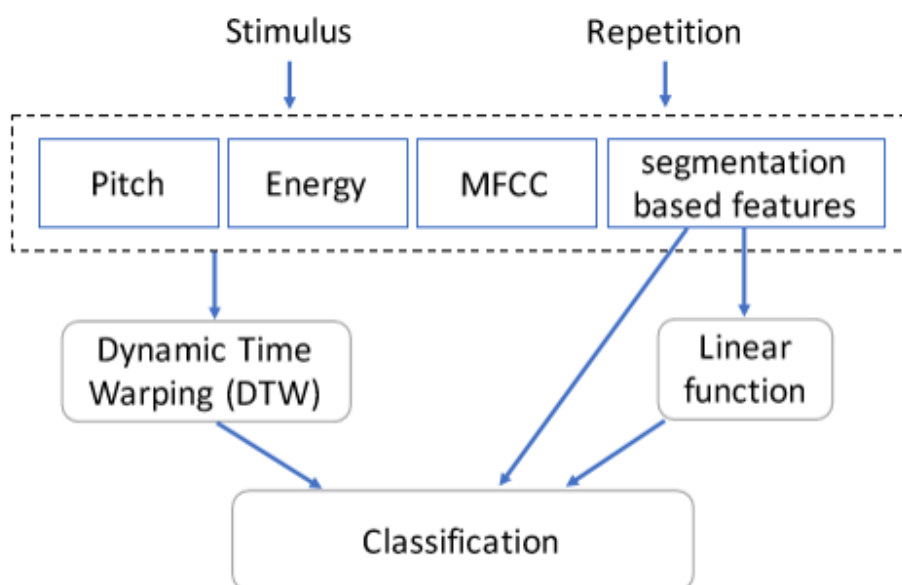


Figure 4.1: Intonation Assessment Method.

4.0.1 Data Collection

Since we could not find a database with the desired characteristics, neither a corpus with autistic children, we had to build our own database. First of all, we asked a European Portuguese (EP) female

speaker to record a total of 20 stimuli (shown in 4.1). Afterwards, we asked several subjects to imitate those stimuli, ending up with a total of 10 participants: 9 healthy adults (3 male and 6 female) and 1 healthy child, leading to a total of 200 recorded utterances. All the recorded utterances were calibrated to a sample rate of 16000Hz, mono channel. Each of the utterances was labelled with 'G', if it was a good imitation, or 'B', if it was a bad imitation, by a non-expert annotator. An intermediate label was initially considered but discarded, given the limited size of the database and the existence of only one annotator. Each subset of 12 utterances (for each of the 20 stimuli) was randomly and equally divided into two distinct folders, one for training our algorithm, and another one for testing it.

Appendix A shows the complete dataset, discriminating between training, and test subsets, and between good or bad labels.

Table 4.1: Stimuli database

Stimuli	Intonations
Banana	Affirmation, Question, Pleasure, Displeasure
Bolo	Affirmation, Question, Pleasure, Displeasure
Gelado	Affirmation, Question, Pleasure, Displeasure
Leite	Affirmation, Question, Pleasure, Displeasure
Ovo	Affirmation, Question, Pleasure, Displeasure

4.1 Feature Extraction

In accordance with several studies on automatic intonation recognition we extracted prosodic features from fundamental frequency (pitch) and the energy contour of the speech signal. In addition, we extracted spectral characteristics in 12 sub-bands derived from the MFCCs, and a set of temporal characteristics coming from pseudo-syllable features extraction script. In this section we will describe how we extracted each one of the used features.

4.1.1 Pitch

Pitch is a subjective auditory sensation and depends on the frequency, the harmonic content, and to a lesser extent on the loudness of a sound. For the pitch contour extraction a library to label music and sounds available for *python* as a free software, named *Aubio*, was used. An example of a pitch contour using this library is shown in figure 4.2.

Pitch Normalization

Two speakers can produce intonationally equivalent utterances even though they have different personal pitch range. For instance, a man and a woman typically result in different pitch 'spans' when plotted on a Hertz scale. Alternative, psycho-acoustic, scales are available, such as semitones, mels, Bark and ERB-rate. Accordingly to the results of the experiment described in [55], the best scale to use is the semitones, since it preserves the main shape of male, female and even children intonation

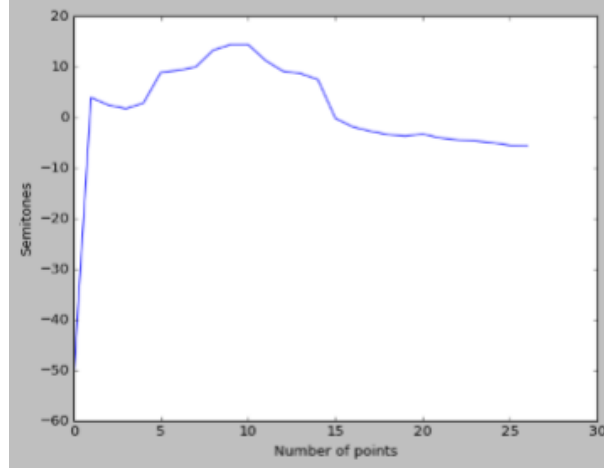


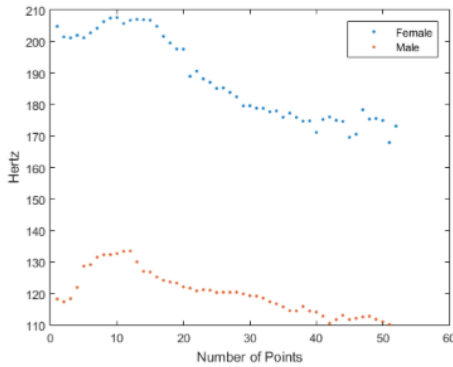
Figure 4.2: Pitch extracted using *Aubio* library.

contours. So, for pitch normalization we convert the speech utterances in a semitone scale, computed by the following equation:

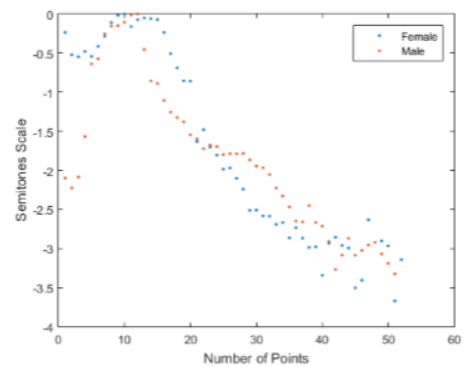
$$ST = \frac{12}{\log 2} \log \frac{x}{100} \quad (4.1)$$

where x is the frequency in Hz that we want to convert, ST is the converted value, and 100 is the reference value for which $ST=0$.

In figure 4.3, we can see two different representations of two pitch contours, for the word 'banana', for male and female speakers. As it is possible to observe, in 4.3(a) the pitch contours are represented in a Hertz scale, where both contours have different ranges and, consequently, we cannot compare them. In 4.3(b) both pitch contours are normalized in a semitones scale, and with this normalization it is possible to compare them.



(a) Pitch contour in Hertz scale.



(b) Pitch contour in Semitones scale.

Figure 4.3: Pitch contour represented in two different scales.

4.1.2 Power

The power plot of a sound is the rate at which sound energy is emitted, reflected, transmitted or received, per unit time. For our database we extracted the power plot using *WaveSurfer*, and an example

is represented in figure 4.4.

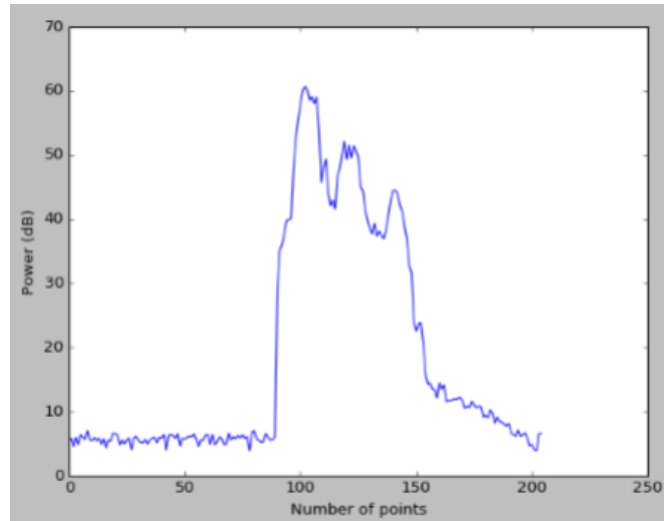


Figure 4.4: Power representation.

4.1.3 MFCCs Extraction

The mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than at high frequencies. Incorporating this scale makes our features match more closely what humans hear. MFCCs are coefficients that together make up a Mel-frequency Cepstrum (MFC), and are derived from a type of cepstral representation of the audio file. The conversion from frequency to mel scale is done by the following equation:

$$M(f) = 1125 \ln \frac{f}{700} \quad (4.2)$$

From an audio file, MFCCs are commonly obtained as:

- Make the Fourier Transform of the signal;
- Map the powers of the spectrum obtained above into the mel scale, by using triangular overlapping windows;
- Take the logarithms of the powers at each mel frequencies;
- Take the discrete cosine transform of the list of mel logarithm powers;
- The MFCCs are the amplitudes of the resulting spectrum.

For our method, we extracted 12 MFCCs for each sound, using *librosa*, which is a python package for music and audio analysis [56]. The way we used it to extract MFCCs features is the one presented in figure 4.5. The obtained MFCCs are then represented in a graph, and we can see an example in figure 4.6.

```

y0, sr0 = librosa.load('./data/banana_e01_u00.wav') #Loading audio files
mfcc0 = librosa.feature.mfcc(y0, sr0, n_mfcc=12)    #Computing MFCC values
librosa.display.specshow(mfcc0)

```

Figure 4.5: How to extract MFCCs features.

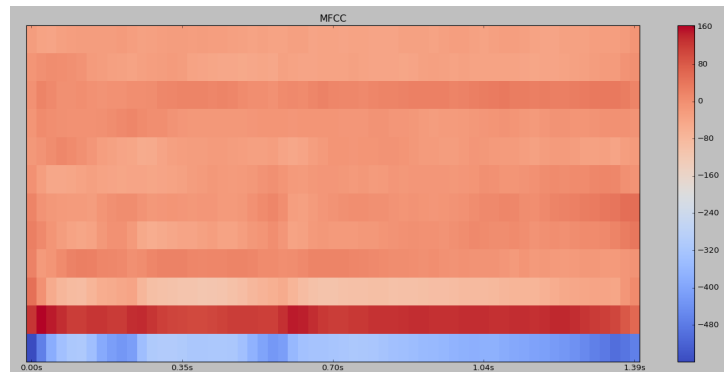


Figure 4.6: Visualization of MFCCs series.

4.1.4 Pseudo-syllables Extraction

For the pseudo-syllable extraction we used a script written in the software program *Praat*, that detects syllable nuclei automatically in order to calculate the speech rate [57]. Locating these syllable nuclei allows for a computation of number of syllables, which is the main objective for the present sub-section. Following, we describe the sequence of steps that the script completes to find syllable nuclei using intensity (dB) and voicedness.

Step 1: The intensity is extracted, with the parameter "minimum pitch" set to 50 Hz, using autocorrelation.

Step 2: All peaks above a certain threshold in intensity to be potential syllables are considered. The threshold was set to 0 or 2 dB above the median intensity measured over the total sound file (0 dB for not filtered sound and 2 dB for filtered sound). They considered the median instead of the mean for the threshold calculation in order to avoid including extreme peaks.

Step 3: They inspected the preceding dip in intensity and considered only a peak with a preceding dip of at least 2 or 4 dB with respect to the current peak as a potential syllable (2 dB for not filtered sound and 4 dB for filtered sound).

Step 4: The pitch contour was extracted, this time using a window size of 100 msec and 20-msec time steps, and all peaks that are unvoiced were excluded.

Step 5: The remaining peaks were considered syllable nuclei and are saved in a TextGrid.

After running the script a window open (figure 4.7), allowing the modification of the default values, having in mind that the higher the silence threshold number (e.g. -30, -40), the lower the chance of finding silent pauses, the higher the minimum dip between peaks the fewer syllables will be found, and the higher the minimum pause duration, the fewer pauses will be found.

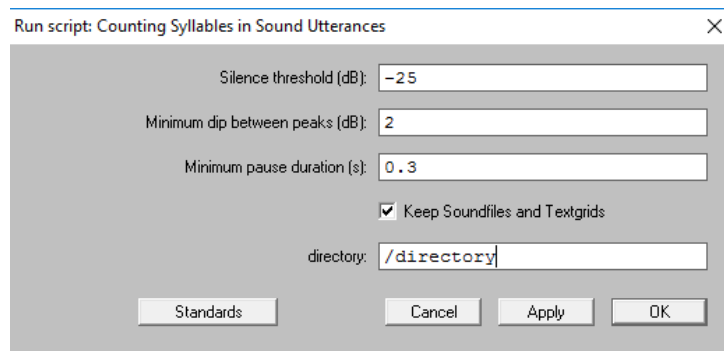


Figure 4.7: Window that allow us to modify some values.

Figure 4.8 shows a sound file of the word "banana" together with the output as TextGrid produced by the script.

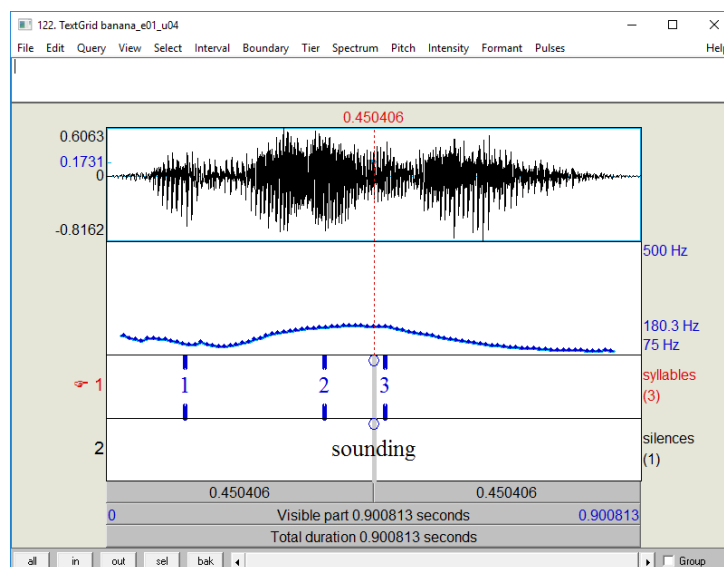


Figure 4.8: TextGrid made by the script.

The script yields several features such as number of syllables, number of pauses, duration (in seconds), phonation time (in seconds), speech rate (nsyll/duration), articulation rate (nsyll/phonation time), and average speaking rate (ASR) (speaking rate/nsyll). These features are displayed in an info window, such as the one in figure 4.9, and we can use and manipulate all the information. Some of these features are not relevant at all for the current scenario (for example, the number of pauses is always zero, showing that the utterances consists of one word).

Praat Info

soundname	nsyll	npause	dur (s)	phonationtime (s)	speechrate (nsyll/dur)	articulation rate (nsyll / phonationtime)	ASD (speakingtime/nsyll)
banana_e01_u00	7	0	1.40	1.40	5.02	5.02	0.199
banana_e01_u01	5	0	1.80	1.80	2.78	2.78	0.360
banana_e01_u02	3	0	1.00	1.00	3.01	3.01	0.333
banana_e01_u03	6	0	1.43	1.43	4.19	4.19	0.238

Figure 4.9: Example of info window produced by the *Praat* script.

4.2 Dynamic Time Warping

Dynamic Time Warping (DTW) is a well known technique to find an optimal distance between two given time-dependent sequences under certain restrictions. This algorithm is normally used for measuring similarity between two time series which may vary in time or speed. Reviewing DTW [58], suppose we have two time series Q and C, of length n and m respectively, where:

$$Q = q_1, q_2, \dots, q_i, \dots, q_n \quad (4.3)$$

$$C = c_1, c_2, \dots, c_i, \dots, c_m \quad (4.4)$$

To align two sequences using DTW, an n-by-m matrix is constructed, where the element of the matrix contains the distance $d(q_i, c_i)$ between the two points q_i and c_j (i.e. $d(q_i, c_i) = (q_i - c_i)^2$). Each matrix element (i,j) corresponds to the alignment between the points q_i and c_i . A warping path W, is a contiguous (in the sense stated below) set of matrix elements that defines a mapping between Q and C. The k^{th} element of W is defined as $w_k = (i,j)_k$ so we have:

$$W = w_1, w_2, \dots, w_k, \dots, w_K, \max(m, n) \leq K < m+n - 1 \quad (4.5)$$

All the elements of DTW are illustrated in figure 4.10.

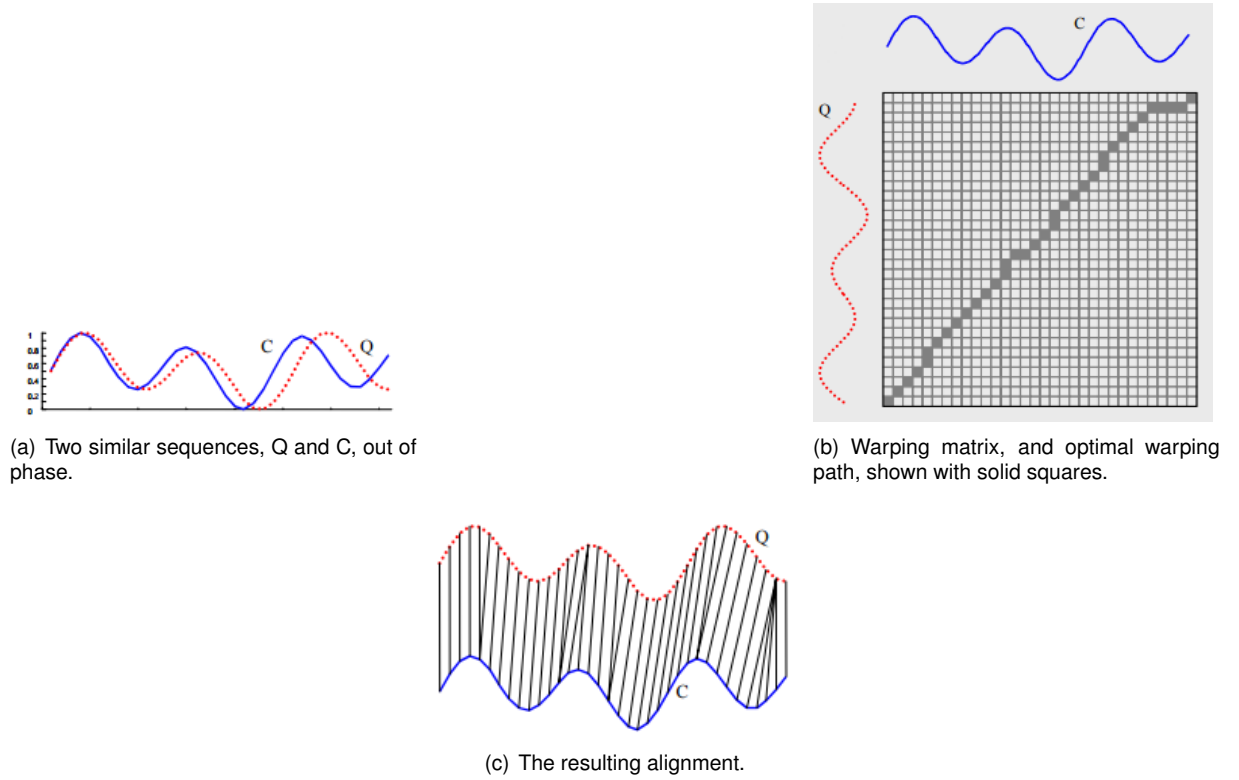


Figure 4.10: DTW graphics representation [58].

In order to achieve the optimal warping path is subjected to several constrains:

- Boundary conditions: $w_1 = (1,1)$ and $w_k = (m,n)$, this requires the warping path to start and finish in diagonally opposite corner cells of the matrix.
- Continuity: Given $w_k = (a,b)$ then $w_{k-1} = (a',b')$ where $a-a' \leq 0$ and $b-b' \leq 0$. This restricts the allowable steps in the warping path to adjacent cells (including diagonally adjacent cells).
- Monotonicity: Given $w_k = (a,b)$ then $w_{k-1} = (a',b')$ where $a-a' \geq 0$ and $b-b' \geq 0$. This forces the points in W to be monotonically spaced in time.

There are exponentially many warping paths that satisfy the above conditions, however we are only interested in the path that minimizes the warping cost:

$$DTW(Q, C) = \min \left\{ \sqrt{\sum_{k=1}^K w_k} \right. \quad (4.6)$$

Besides, DTW allow us to compute the distance function that will give the final cost between the comparison of two signals. If Q and C are both K -dimensional signals, then metric prescribes $d_{mn}(Q,C)$, the distance between the m_{th} sample of Q and the n_{th} sample of C .

Our DTW module was an existent *python* module, that give us the optimal warping path, as we can see in figure 4.11, and a cost that allow us to compute the threshold for which the intonation imitation is correct.

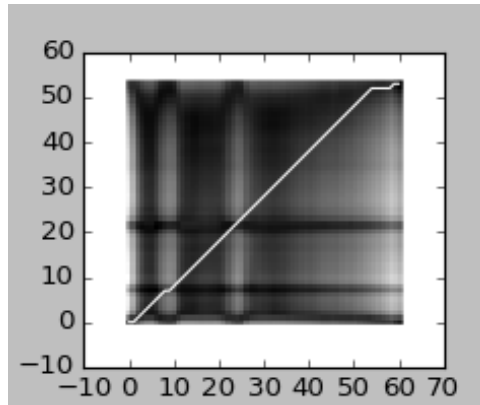


Figure 4.11: Representation of the optimal warping path.

The DTW module was applied to pitch, power and MFCCs. Pseudo-syllables may be directly compared.

4.3 Classification

The classification algorithm consists in measuring the distance between all points of a stimulus with all points of its imitations. Applied the algorithm to all imitations of the training class, we were able to obtain the distance between all the stimuli and its imitations, as shown in figure 4.12. Afterwards, the mean and the standard deviation of the distances of the imitations classified as good, and for the ones classified as bad were obtained, giving the cost for good and bad imitations for that feature.

```

NDTW (./new_data_train\banana_e01_u00.wav , ./new_data_train\banana_e01_u01_g.wav) (g): 3.456475
NDTW (./new_data_train\banana_e01_u00.wav , ./new_data_train\banana_e01_u03_g.wav) (g): 2.238859
NDTW (./new_data_train\banana_e01_u00.wav , ./new_data_train\banana_e01_u05_B.wav) (B): 11.642419
NDTW (./new_data_train\banana_e01_u00.wav , ./new_data_train\banana_e01_u07_B.wav) (B): 14.244510
NDTW (./new_data_train\banana_e01_u00.wav , ./new_data_train\banana_e01_u09_G.wav) (G): 1.957473

```

Figure 4.12: Cost between a stimulus and its imitation.

For setting a threshold, we calculated the mean between both good and bad costs. In figure 4.13 the centroids obtained with the training class are presented. Having the threshold defined, we tested our method with the test class of the database, evaluating if the cost for each imitation is under the defined threshold, classifying it as a 'C' (good imitation), or above the threshold, classifying it as 'I' (bad imitation), as illustrated in the example of figure 4.14.

```

Centroids {'B': 12.94346437507715, 'G': 2.9747990824347395} threshold: 7.95913172876
Standard Deviation {'B': 1.3010451927399025, 'G': 1.2589647775175625}

```

Figure 4.13: Set of informations given by the train class.

```

NDTW (./new_data_test\banana_e00_u00.wav , ./new_data_test\banana_e00_u01_G.wav) (G): 4.143627 (C)
NDTW (./new_data_test\banana_e00_u00.wav , ./new_data_test\banana_e00_u03_G.wav) (G): 3.965048 (C)
NDTW (./new_data_test\banana_e00_u00.wav , ./new_data_test\banana_e00_u05_g.wav) (g): 3.846806 (C)
NDTW (./new_data_test\banana_e00_u00.wav , ./new_data_test\banana_e00_u09_G.wav) (G): 3.794850 (C)

```

Figure 4.14: Testing the method.

For computing the costs of pseudo-syllable features, we directly compared them, by defining a distance function. The implemented distance function is given by:

$$D = \sum_{i=1}^N a \cdot |x_i - y_i| \quad (4.7)$$

where a is the multiplicative factor, x_i is the feature corresponding to the stimuli, y_i is the feature corresponding to an imitation and D is the cost of this comparison. The classification method for achieving the threshold was the same used with other features.

A decision tree classifier was also used, thus allowing to perform a classification not only based on one feature but also based on the combination of several features. The decision tree was trained using the existing training data, and it was restricted to a given maximum depth thus restricting the number of decisions performed.

4.4 Tests and Results

In this section, we present the results of evaluating the developed method, applied separately to each set of features. Once the threshold was tuned, with the data of the training set, we obtained the correspondents "correct" (C) or "incorrect" (I) labels, for each utterance of the test set. We then computed a performance measure of the algorithm, the accuracy. The accuracy measure was the total

Table 4.2: Results obtained with MFCCs.

MFCCs	
Distance Function	Accuracy
$\sqrt{\sum_{i=1}^N (x_i - y_i)^2}$	65.5%
$\frac{\sum_{i=1}^N x_i \cdot y_i}{\sqrt{\sum_{i=1}^N x_i^2} \cdot \sqrt{\sum_{i=1}^N y_i^2}}$	81.1%
$\frac{\frac{1}{N} \sum_{i=1}^N x_i \cdot y_i - \mu_X \cdot \mu_Y}{\sigma_X \cdot \sigma_Y}$	83.3%

Table 4.3: Results obtained with Pitch.

Pitch	
Distance Function	Accuracy
$\ x - y\ $	77.5%
$ x - y $	77.5%
$ x^2 - y^2 $	78.8%

of correct classifications, which is the percentage of cases where the algorithm correctly classified the utterances.

The first tests, involved the spectral characteristics in 12 sub-bands derived from the MFCCs. In order to achieve the best results possible, we applied several distance functions. In table 4.2 we present the results of the algorithm performed with MFCCs. As we can see, the best result was performed when the distance was calculated with correlation, obtaining an accuracy of 83.3%. The algorithm was also performed with other distance functions in addition to those presented in the table, but the results were not good.

In table 4.3, the results of the performance of the algorithm using pitch are presented. The best accuracy obtained was 78.8%, with the distance function $|x^2 - y^2|$.

The results of the performance of the algorithm using power are presented in table 4.4. The best accuracy obtained was 71.4%, with the distance function $|x^2 - y^2|$, similarly to what happened when using pitch.

In table 4.5, we present the results obtained using the set of features coming from pseudo-syllables, using different cost functions, where we varied the multiplicative factor of some features. For function D1 we attributed the same multiplicative factor to all features and, as we were already expecting, the obtained accuracy was not good. The function with best accuracy results was D2, with an accuracy of 76.7%.

Summing up, the best accuracies when applying the classification algorithm based on a threshold for each extracted feature, and also for the fusion of all features, using the decision tree, are presented

Table 4.4: Results obtained with Power.

Power	
Distance Function	Accuracy
$\ x - y\ $	70%
$ x - y $	70%
$ x^2 - y^2 $	71.4%

Table 4.5: Results obtained with Pseudo-syllables features.

Pseudo-syllables features	
Distance Function	Accuracy
D1	55.6%
D2	76.7%
D3	61.1%

in table 4.6. Concluding, the best accuracy was verified using MFCCs. An important result was the one computed by the fusion of features, since it allow us to obtain important conclusions.

Table 4.6: Final Results.

Feature		Accuracy	
		Mean&stdev	Decision Tree
Framed-based DTW	MFCCs	83.3%	82.2%
	Pitch	72.2%	72.2%
	Energy	70.0%	74.4%
	Fusion	—	77.8%
Segment-based	Pseudo-syllable features	—	73.3%
Fusion		—	75.5%

4.5 Summary

In the present chapter, we described our approach to build an intonation assessment method and also the results of evaluating this method.

First of all we created our own database, which is constituted by 20 stimuli utterances and 240 imitations utterances, recorded by EP speakers. In order to compute an automatic classification method, all the utterances were labelled with 'G', for good imitations, and with 'B', for bad imitations. After this task, we extracted all the desired features and built a very simple threshold-based classifier. The threshold was defined for each feature set using the training data. The classifier was then applied to the test data to label the intonation as correct or incorrect, yielding an accuracy. In order to know what feature is more informative in terms of intonation imitation, the algorithm was applied separately to each feature.

The obtained results show that the highest accuracy (83.3%) was achieved using MFCCs, but pitch, energy, and pseudo-syllables also proved to be informative. The obtained results for the fusion of the framed-based features was 77.8%, and the accuracy results for the fusion including also the segment-based features was 75.5%. In both fusion results energy is the first selected feature in the decision tree and energy is already covered in MFCCs, therefore the later are very robust in this task, being the one with the best performance, even better than fusion.

Chapter 5

New Contributions to the Virtual Therapist

As previously referred, it is our intention to extend the VITHEA-Kids, by adding a set of prosodic exercises. The first step towards our goal is the development of a stand-alone mobile application.

In this thesis we developed an android application for children diagnosed with ASD, composed by a set of exercises, whose main objective is to improve and acquire prosody skills, that are important not only for educational purposes, but also for leisure, so that children can share their preferences with others, comment on existing contents, and communicate with their peers more expressively. Furthermore, it is an extension of therapy sessions, in home environment, where children feel more relaxed.

In this chapter we document how the application was designed and what exercises do we suggest for the purpose of improving the prosodic skills. In 5.1 we present some requirements that we took into consideration when developing the APP, that meet the users needs. As for sections 5.2 and 5.3, we present the recorded stimuli and the prompt system implemented. In 5.4 we describe the development and implementation itself, explaining each exercise evolved, from the initial prototype to the present version, and also describe the decisions that were made throughout this process, why we added new features and based on what groundings.

5.1 Requirements Analysis and Definition

This section will define what were the requirements for this APP, what should it do and what features should it provide in order to reflect the needs of its users. These requirements reflect the objectives of this thesis, in a perspective oriented to development. There are two types of requirements, functional requirements and non-functional requirements and to define them we made a lot of research and we received some therapist feedback.

5.1.1 Functional Requirements

Concerning with functional requirements, it reflects how the system should react, behave and what should it provide given a certain condition. For the overall functional requirements of this APP, we set the following list:

- The APP should have different types of prosodic exercises.
- The APP should give punctuation for each correct exercise.
- The user should be able to choose between different types of exercises.
- The user should be able to customize each exercise.
- The user should be able to finish the exercise any time.
- The APP should have reinforcements.
- The APP should have a prompt system.

5.1.2 Non-Functional Requirements

Non-functional requirements are not directly connected to the services delivered to the user but on which such services depend to better perform their role. These kind of requirements are related to system properties, such as reliability and response time, and affect the overall architecture of a system. Having this in mind we define the following requirements:

- The APP should have a clean interface.
- The navigation between scenes should be easy and fast.
- The APP should have an intuitive interface.
- The APP should not have words/sentences written.

5.2 Recorded Stimuli

In order to have a correct selection of the stimuli to be used in certain exercises, it is necessary to take into consideration a range of factors, namely psycholinguistic indexes, such as the age of acquisition of Portuguese words [59]. Another important aspect to have in mind while select the correct stimuli is the syllabic extension (no more than three syllables), frequency, easy representation, and the age of acquisition should be less or equal to five years old (for our particular audience). Since it is a topic of easy representation and comprehension, it was decided to use stimuli corresponding to the food category. After selecting the correct stimuli, we separate them into three lists, each one for a corresponding task: a list for the intonation distinction task (table 5.1), a list for the affection recognition task (table 5.2), and another one for the imitation task (table 5.3). While organizing the lists we had into consideration not repeating the same intonation more than three times in a row.

Table 5.1: Recorded stimuli for the intonation distinction task.

Intonation Distinction Task		
Order	Stimuli	Category
1	Manteiga	Question - Question
2	Limão	Displeasure - Affirmation
3	Papa	Pleasure - Question
4	Mel	Displeasure - Pleasure
5	Massa	Question - Affirmation
6	Pão	Affirmation - Affirmation
7	Couve	Displeasure - Affirmation

Table 5.2: Recorded stimuli for the affection recognition task.

Affect Recognition Task		
Order	Stimuli	Category
1	Couve	Displeasure
2	Morangos	Pleasure
3	Papa	Pleasure
4	Limão	Displeasure
5	Batata	Pleasure
6	Bolacha	Pleasure
7	Brócolos	Displeasure

Table 5.3: Recorded stimuli for the imitation task.

Imitation Task		
Order	Stimuli	Category
1	Limão	Affirmation
2	Batata	Pleasure
3	Massa	Question
4	Couve	Displeasure
5	Bolacha	Pleasure
6	Pão	Affirmation
7	Manteiga	Question

5.3 Recorded Utterances

The prompt system is a set of cues with the aim of helping the player in a certain task. In [60] it is proved that graded cueing has good results and is well suited for most children with ASD. The generalized prompts for the application are:

- P0, main menu - The agent explains how to choose a game: "Choose the game that you want to play."
- P1, general - The agent gives general indications: "Press the sound button, to hear the explanations."
- P2, general - The agent gives general indications: "If you want to return to the main menu, press the button home."
- P3, general - The agent gives general indications: "Listen to the sound by pressing the musical notes button."

As for positive reinforcement, where the agent encourages the user to continue with the good work, they are the following:

- P0 - "Good!"
- P1 - "You got it right."
- P2 - "You made it."
- P3 - "Very good!"
- P4 - "You are almost done."
- P5 - "Uau!"
- P6 - "Amazing!"
- P7 - "You are going very well."

Regarding negative reinforcement, where the agent encourages the user to continue the game despite the answer being wrong, they are:

- P0 - "I think that is not the correct answer."
- P1 - "That is not correct."
- P2 - "Oh, no!"
- P3 - "Try one more time."
- P4 - "Ups, it is not like that."
- P5 - "Ooooooh!"

Besides, for each exercise, we recorded some specific utterances, like the exercise explanation, since it is different for each exercise. Below, we expose the recorded prompts for each exercise:

- P0 - The agent gives indications about the first exercise, first version: "Are the two images the same?"
- P1 - The agent gives indications about the first exercise, second version: "Press the different image."
- P2 - The agent gives indications about the second exercise: "Are the two sounds equal?"
- P3 - The agent gives indications about the third exercise: "Imitate what is said."
- P4 - The agent gives indications about the fourth exercise: "Listen well the food item and choose whether the person likes it or not."
- P5 - The agent gives indications about the fifth exercise: "Press the screen to hear a high sound."

- P6 - The agent gives indications about the fifth exercise: "Press the screen to hear a low sound."
- P7 - The agent gives indications about the fifth exercise: "Listen the sound, is it high or low?"

A full list of the recorded prompts, as well as the recorded stimuli, in Portuguese, is available in appendix B.

5.4 New Exercises

In a generalized way, the structure of the APP is the one presented in figure 5.1. In the main menu there are five buttons, where the user can click, corresponding to each one of the five exercises available. Since the audience are children diagnosed with ASD, that can not read, the identification of each exercise is made by an icon. The choice of the games was based on a set of studied surveys, referred in Chapter 2 and Chapter 3. With this set of exercises we pretend to develop the reception and processing of sound skills as well as the imitation of stimulus related with the most basic level of phonetic processing, in which meaning is not involved. Besides, it is our intention to develop the capacity to understand and express prosody to display the affective, pragmatic, grammatical and interactive functions.

The development of the application was made in an integrated development environment for the android platform, the Android Studio.

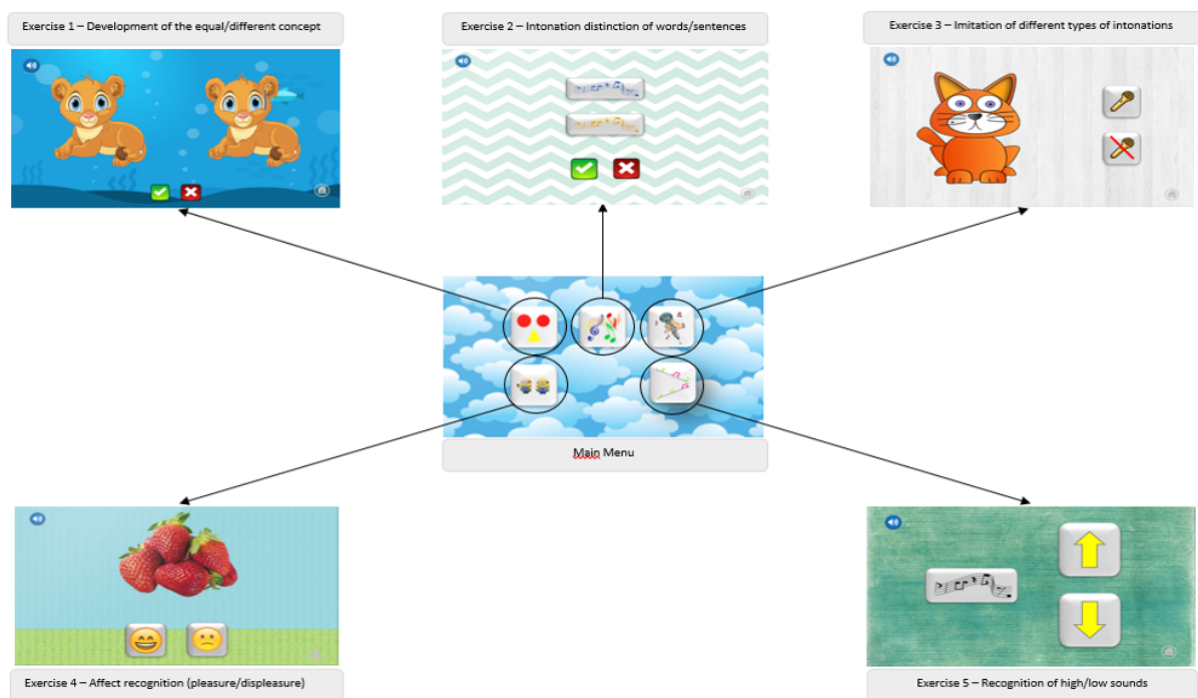


Figure 5.1: Implemented Structure.

The focus of our application is the development of prosodic exercises, however we decided to integrate an exercise that establishes the connection between our APP and VITHEA-Kids. The exercise has the objective of developing and stimulating the equal/different concept. This is a simple task for implementation, as we can see in figure 5.2, but it is very important for children with ASD to understand

this concept, and fundamental for other exercises. There are two different versions of this game. In the first version the child should analyse two images displayed on the screen and click the check button if the images are equal or the wrong button if the images are different, as shown in figure 5.3(a). For the second version we display three images in the screen and the child should click in the different image, as shown in figure 5.3(b).

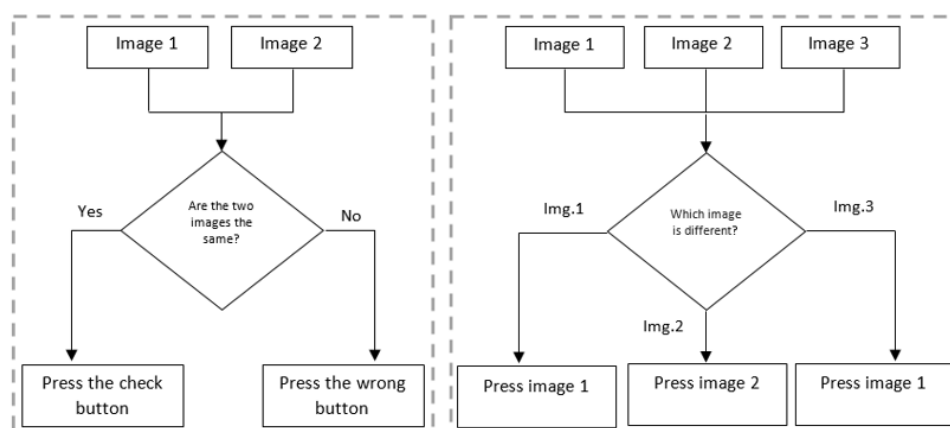
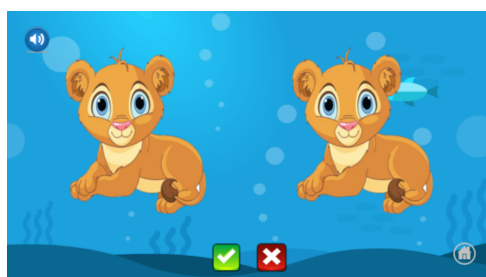


Figure 5.2: Architecture of the equal/different concept exercise.



(a) Version 1

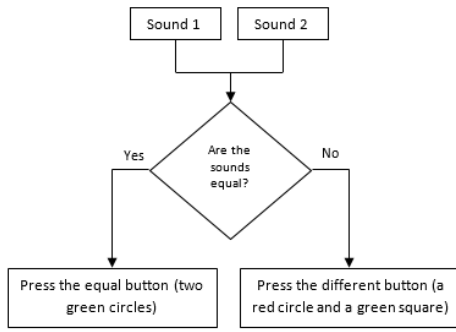


(b) Version 2

Figure 5.3: Layout of the equal/different concept exercise.

5.4.1 Intonation Distinction

The second exercise is about intonation distinction of words. The objective of this exercise is develop the skills of an ASD children to understand intonation changes in short stimulus (words). For this task the discrimination paradigm of "equal *versus* different" is used and the procedure consists in presenting two sound stimuli without any segmental information. After hearing the two stimuli, the user only has to understand whether the sounds are equal, and choose the check button or different and choose the wrong button, as referred in image 5.4(a). The layout of the exercise is presented in figure 5.4(b). For this task in particular, the different intonations are affirmation/question and pleasure/displeasure.



(a) Architecture of the intonation distinction exercise.



(b) Layout of the intonation distinction exercise.

Figure 5.4: Architecture and layout of the intonation distinction exercise.

5.4.2 Intonation Imitation

The third exercise objective is to develop the children skills to imitate different types of intonations in short stimuli, composed by one word. This exercise has integrated the intonation assessment method, since we pretend to evaluate if the children made a good imitation of the stimuli or not. This exercise is extremely important because it will allow children to have more confidence when expressing themselves with emotion or to express their tastes while interacting with someone. The architecture of the present exercise is shown in figure 5.5. In order to make this task as attractive as possible, we design a cute kitty, that moves while speaking. In figure 5.6 is represented the layout that is shown in the APP.

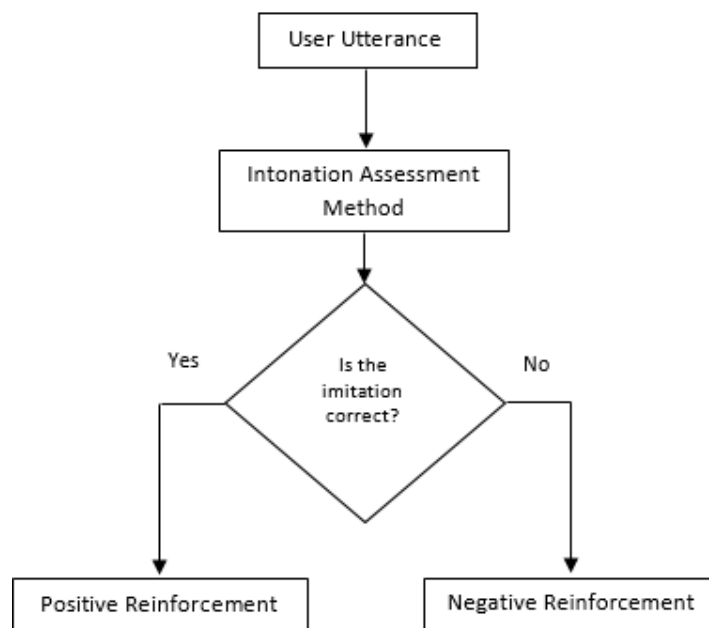


Figure 5.5: Architecture of the intonation imitation exercise.

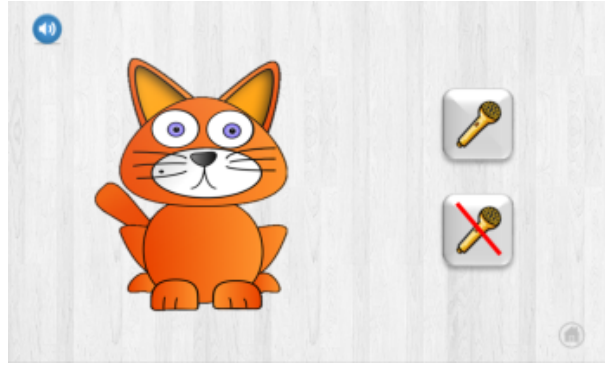


Figure 5.6: Layout of the intonation imitation exercise.

5.4.3 Affect Recognition

This exercise, which represents an affection task, is concerned with the understanding and use of prosody to express pleasure/displeasure. This exercise intention is to evaluate and develop the receptive component of the affection task. For this task, we implemented the architecture of figure 5.7. As for the way that the game works, a food item appears on the screen, followed by an auditory stimulus, namely the food item name pronounced with pleasure/displeasure. The answer consists in select one of two buttons that appear on the screen simultaneously, a button with a smiley face in case the user consider the stimulus corresponding to pleasure, and a sad face in case the user consider the stimulus corresponding to displeasure. In figure 5.8 the layout of the present exercise is represented.

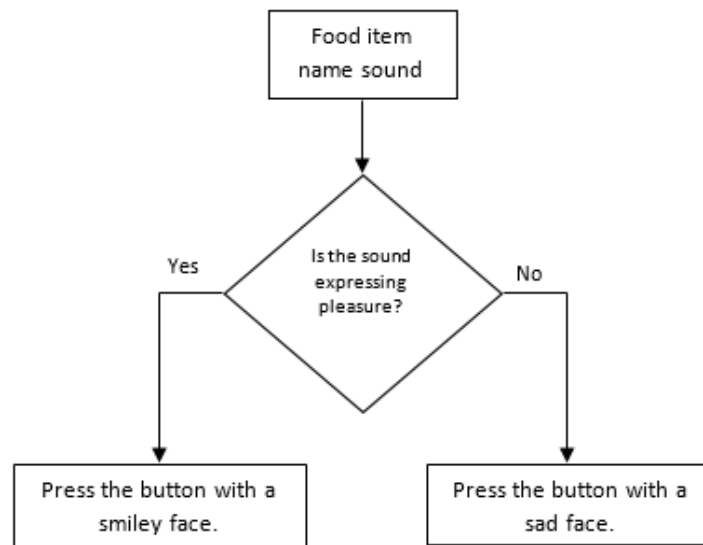


Figure 5.7: Architecture of the affect recognition exercise.

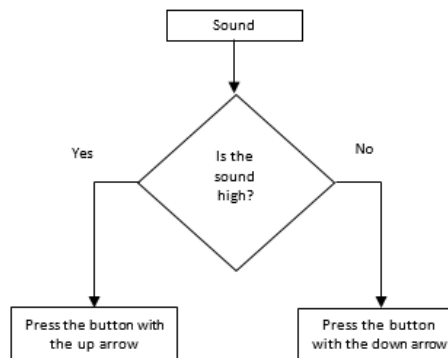
5.4.4 Up/Down Recognition

The conception of the present game was inspired in a study made by Thorson et al. (2016) [43], mentioned in 3.1.2. For our APP we made some adaptations, always focus on developing the capacity of the children with ASD of distinguish low and high sounds. The architecture is the one present in figure

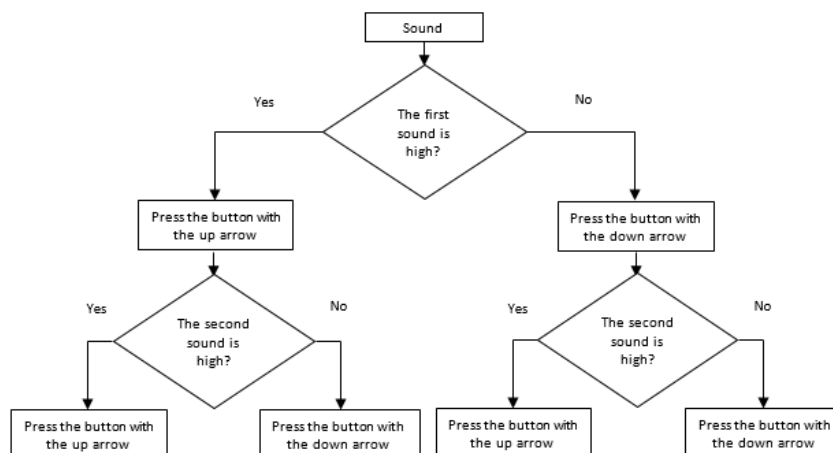


Figure 5.8: Layout of the affect recognition exercise.

5.9 and there are two types of exercises for the present game. The first exercises consist on listen a single sound (starting with animal sounds before proceeding to human sounds) and then press the up arrow for high sounds or the down arrow for low sounds, such in figure 5.9(a). The next version is a little more complex since a sequence of two sounds is displayed and then the user has to press the arrows in accordance with the sounds (for example, if the sequence is high-high, the user needs to press two times the button with the up arrow), such in figure 5.9(b). The layout of this exercise is quite simple, as shown in figure 5.10. In order for children to better understand this exercise, we will give, at the beginning, an example of a high and a low sound.



(a) Single sound.



(b) Sequence of sounds.

Figure 5.9: Architecture of the high/low recognition exercise.

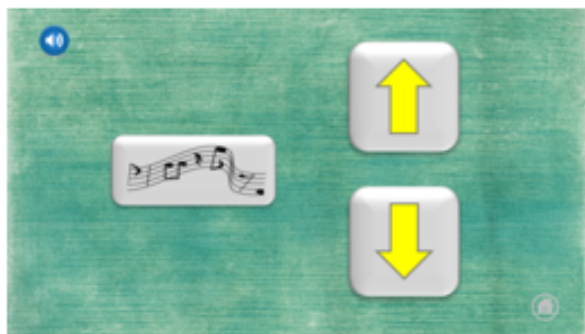


Figure 5.10: Layout of the high/low recognition exercise.

5.5 Summary

In this chapter we presented some developed exercises, for an Android application, which will be integrated later in the Virtual Therapist, VITHEA-KIDS. The developed APP respects all the requirements initial proposed, being super easy to manipulate and understand. Unfortunately, and because the lack of time, since it depends on the availability of Hospital Garcia da Horta, we were not able to evaluate the application with children diagnosed with ASD. However, the idea of the conception of prosodic exercises emerged from a necessity exposed by psychologists and therapists, that daily work with such children. All the presented exercises were carefully chosen, after studying the lack of communication and prosodic skills and their needs in this field.

As previously mentioned, the development of the prosodic skills is extremely important, since it will allow children to understand several dynamics while communicating with someone, as well as expressing themselves with more intonation, helping them being more confident and making the communication more fluent and understandable.

We decided to build this application in Android environment, since it is a free interface for the developer and also because it is more accessible for the public in general

Chapter 6

Conclusions and Future Work

Having into consideration all the goals of this thesis and all the work presented in the previous chapters, in the present chapter it is important to reflect about all the developed work, from the studies realized to the implementation itself, as well as the obtained results. Afterwards, we present some ideas and suggestions for future work on this topic.

6.1 Conclusions

As stated in chapter 1, the main goal of this thesis was to fill a market gap, since there is not available applications for prosodic training for children with ASD, that have a lot of impairments in this field. Prosodic training is very important, as it may helps an autistic child to develop communication with emotion and give more confidence to children to talk and interact with people. The idea of fill this market gap came from therapists, which is a great indicator of the need of having a tool to develop prosodic skills as a therapy complement.

Despite the fact that the original intention was to integrate prosodic exercises in an existing platform for children diagnosed with ASD, the actual implementation was an android application combining a set of prosodic exercises, inspired in several interventions and analysed surveys in chapters 2 and 3. We came out with exercises that aim to develop the capacity to understand and express prosody to display the affective, pragmatic, grammatical and interactive functions. During the development of this application we also had into account some functional and non-functional requirements, for such a specific audience.

One of the implemented exercises has integrated an intonational assessment method, which was another goal of this thesis. In fact, we implemented this assessment method an evaluated the performance of the algorithm separately for each feature set, and also by making a fusion of all features. The performance of the proposed method was evaluated only for healthy subjects, yielding accuracy values between 70% and 83.3%, depending on the selected feature.

Unfortunately, we were not able to perform tests with the application itself, since we were dependent on the availability of the hospital, and we were running out of time.

6.2 Future Work

With the end of our study, we found a few aspects that can be approached in subsequent work, as from the game perspective and also in the study of the intonation assessment method. Starting with the study, a long-term experiment should be done with more participants, both health and children with ASD, in order to have a more reliable method and better define the threshold. We also suggest the combination of all available features in an algorithm, for achieving better results.

Regarding the APP itself, we suggest its full integration in the VITHEA-KIDS platform, since it has not exercises related with the development of prosodic skills. In favour of having a more attractive and reliable app, for such a particular audience, we suggest the synchronization of the animated toy with speech. The APP should also have an user interface, where the therapists as well as caregivers could insert images and sounds in accordance with each child preferences, and also monitoring the progress of the child, which represents a concern of the VITHEA-KIDS itself. Finally, it would be interesting to evaluate the performance of the application with autistic children.

Another important aspect to have in mind in the future, is the possibility to the caregiver to adjust the threshold in the imitation task, in accordance with the evolution of the child. This way the caregiver could be more exigent with the child, and the progresses will surely be more notable.

Bibliography

- [1] *Diagnostic and Statistical manual of mental disorders: DSM-5*. American Psychiatric Association, American Psychiatric Association Arlington, VA, 5th edition, 2013.
- [2] C. for Disease Control and Prevention. Autism spectrum disorder - data and statistics, 2014. URL <http://www.cdc.gov/ncbddd/autism/data.html>. [Online; accessed 30-December-2015].
- [3] A. J. Baxter, T. Brugha, H. Erskine, R. Scheurer, T. Vos, and J. Scott. The epidemiology and global burden of autism spectrum disorders. *Psychological medicine*, 45(03):601–613, 2015.
- [4] M. Elsabbagh, G. Divan, Y.-J. Koh, Y. S. Kim, S. Kauchali, C. Marcín, C. Montiel-Nava, V. Patel, C. S. Paula, C. Wang, et al. Global prevalence of autism and other pervasive developmental disorders. *Autism Research*, 5(3):160–179, 2012.
- [5] G. Oliveira. *Edipemiologia do autismo em Portugal*. PhD thesis, Faculdade de Medicina da Universidade de Coimbra, 2005.
- [6] N. Naoi. Intervention and treatment methods for children with autism spectrum disorders. In J. L. Matson, editor, *Applied Behavior Analysis for Children with Autism Spectrum Disorders*, chapter 4, pages 67–81. Springer, 2009.
- [7] S. Shamsuddin, H. Yussof, L. Ismail, F. A. Hanapiah, S. Mohamed, H. A. Piah, and N. I. Zahari. Initial response of autistic children in human-robot interaction therapy with humanoid robot nao. In *Signal Processing and its Applications (CSPA), 2012 IEEE 8th International Colloquium on*, pages 188–193. IEEE, 2012.
- [8] B. Scassellati, H. Admoni, and M. Mataric. Robots for use in autism research. *Annual review of biomedical engineering*, 14:275–294, 2012.
- [9] N. Giullian, D. Ricks, A. Atherton, M. Colton, M. Goodrich, and B. Brinton. Detailed requirements for robots in autism therapy. In *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*, pages 2595–2602. IEEE, 2010.
- [10] A. Duquette, F. Michaud, and H. Mercier. Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Autonomous Robots*, 24(2):147–157, 2008.

- [11] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard. Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, 4(2):105–120, 2005.
- [12] L. Kanner. Autistic disturbances of affective contact. In *Acta paedopsychiatrica*, volume 35, pages 100–136. 1943.
- [13] H. Asperger. Autistic psychopathy in childhood. In *Autism and Asperger Syndrome*. Cambridge University Press.
- [14] M. L. Bauman and T. L. Kemper. *The Neurobiology of Autism*. The Johns Hopkins University Press, 2nd edition, 2005.
- [15] *Diagnostic and Statistical manual of mental disorders*. American Psychiatric Association, Washington DC: APA, 4th edition, 2000.
- [16] P. K. Kuhl, S. Coffey-Corina, D. Padden, and G. Dawson. Links between social and linguistic processing of speech in preschool children with autism: behavioral and electrophysiological measures. *Developmental science*, 8(1):F1–F12, 2005.
- [17] P. K. Kuhl. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843, 2004.
- [18] T. Owley, L. Walton, J. Salt, S. J. Guter, M. Winnega, B. L. Leventhal, and E. H. Cook. An open-label trial of escitalopram in pervasive developmental disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44(4):343–348, 2005.
- [19] E. Bal, E. Harden, D. Lamb, A. V. Van Hecke, J. W. Denver, and S. W. Porges. Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. *Journal of autism and developmental disorders*, 40(3):358–370, 2010.
- [20] S. Dodd. *Understanding the Autism*. Elsevier, Australia, 2004.
- [21] J. E. Ringdahl, T. Kopelman, and T. S. Falcomata. Applied behavior analysis and its application to autism and autism related disorders. In J. L. Matson, editor, *Applied Behavior Analysis for Children with Autism Spectrum Disorders*, chapter 2, pages 15–32. Springer, 2009.
- [22] J. Coolican, I. M. Smith, and S. E. Bryson. Brief parent training in pivotal response treatment for preschoolers with autism. *Journal of Child Psychology and Psychiatry*, 12:1321–1330, 2010.
- [23] B. F. Skinner. *Verbal Behavior*. Prentice-Hall, Inc, Cambridge, Massachusetts, 1957.
- [24] S. I. Greenspan and S. Wieder. Floortime as a family approach. In *Engaging Autism: Using the Floortime Approach to Help Children Relate, Communicate, and Think*, chapter 13, pages 163–177. Da Capo Press, 2009.
- [25] S. E. Gutstein and R. K. Sheely. *Relationship Development Intervention with children, Adolescents and Adults*. Jessica Kingsley Publishers, United Kingdom, 2002.

- [26] G. B. Mesibov, V. Shea, and E. Schopler. *The TEACCH Approach to Autism Spectrum Disorders*. Springer Science, 2004.
- [27] B. M. Prizant, A. M. Wetherby, E. Rubin, and A. C. Laurent. The scerts model. *Infants and Young Children*, 16(4):296–316, 2003.
- [28] J. McCann and S. Peppé. Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders*, 38(4):325–350, 2003.
- [29] M. Filipe. *Prosodic Abilities in Typically Developing Children and those Diagnosed with Autism Spectrum Disorders - Clinical Implications for Assessment and Intervention*. PhD thesis, University Porto - Faculdade de Psicologia e de Ciências da Educação, 2014.
- [30] G. De Leo and G. Leroy. Smartphones to facilitate communication and improve social skills of children with severe autism spectrum disorder: special education teachers as proxies. In *Proceedings of the 7th international conference on Interaction design and children*, pages 45–48. ACM, 2008.
- [31] R. Jordan and S. Powell. *Understanding and teaching children with autism*. Wiley, 1995.
- [32] M. Davis, N. Otero, K. Dautenhahn, C. L. Nehaniv, and S. D. Powell. Creating a software to promote understanding about narrative in children with autism: Reflecting on the design of feedback and opportunities to reason. In *Development and Learning, 2007. ICDL 2007. IEEE 6th International Conference on*, pages 64–69. IEEE, 2007.
- [33] F. G. Happé. Central coherence and theory of mind in autism: Reading homographs in context. *British journal of developmental psychology*, 15(1):1–12, 1997.
- [34] O. Bogdashina. *A reconstruction of the sensory world of autism*. Sheffield Hallam University Press Sheffield, 2001.
- [35] C. Putnam and L. Chong. Software and technologies designed for people with autism: what do users want? In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, pages 3–10. ACM, 2008.
- [36] M. Dawe. Desperately seeking simplicity: how young adults with cognitive disabilities and their families adopt assistive technologies. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1143–1152. ACM, 2006.
- [37] M. Moore and S. Calvert. Brief report: Vocabulary acquisition for children with autism: Teacher or computer instruction. *Journal of autism and developmental disorders*, 30(4):359–362, 2000.
- [38] A. Bosseler and D. W. Massaro. Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. *Journal of autism and developmental disorders*, 33(6):653–672, 2003.
- [39] D. W. Massaro and A. Bosseler. Read my lips: The importance of the face in a computer-animated tutor for vocabulary learning by children with autism. *Autism*, 10(5):495–510, 2006.

- [40] O. E. Hetzroni and U. Shalem. From logos to orthographic symbols: A multilevel fading computer program for teaching nonverbal children with autism. *Focus on Autism and Other Developmental Disabilities*, 20(4):201–212, 2005.
- [41] A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, and I. P. Martins. Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language*, 27(6):1235–1248, 2013.
- [42] V. Mendonça, L. Coheur, and A. Sardinha. Vithea-kids: a platform for improving language skills of children with autism spectrum disorder. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 345–346. ACM, 2015.
- [43] J. Thorson, S. Meyer, D. Plesa-Skwerer, R. Patel, and H. Tager-Flusberg. Assessing prosody in minimally to nonverbal children with autism. *Speech Prosody 2016*, pages 1206–1210, 2016.
- [44] S. Parsons, P. Mitchell, and A. Leonard. The use and understanding of virtual environments by adolescents with autistic spectrum disorders. *Journal of Autism and Developmental disorders*, 34(4):449–466, 2004.
- [45] A. Leonard, P. Mitchell, and S. Parsons. Finding a place to sit: a preliminary investigation into the effectiveness of virtual environments for social skills training for people with autistic spectrum disorders. *Virtual Reality and Associated Technologies, Veszprem, Hungary, University of Reading*, 2002.
- [46] P. Mitchell, S. Parsons, and A. Leonard. Using virtual environments for teaching social understanding to 6 adolescents with autistic spectrum disorders. *Journal of autism and developmental disorders*, 37(3):589–600, 2007.
- [47] D. F. Cihak, C. C. Smith, A. Cornett, and M. B. Coleman. The use of video modeling with the picture exchange communication system to increase independent communicative initiations in preschoolers with autism and developmental delays. *Focus on Autism and Other Developmental Disabilities*, 27(1):3–11, 2012.
- [48] J. Ohene-Djan. Winkball for schools: An advanced video modelling technology for learning visual and oral communication skills. In *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on*, pages 687–689. IEEE, 2010.
- [49] S. Witt and S. Young. Computer-assisted pronunciation teaching based on automatic speech recognition. *Language Teaching and Language Technology Groningen, The Netherlands*, 1997.
- [50] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt. Automatic detection of phone-level mispronunciation for language learning. In *EUROSPEECH*, 1999.
- [51] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari. The sri eduspeaktm system: Recognition and pronunciation scoring for language learning. *Proceedings of InSTILL 2000*, pages 123–128, 2000.

- [52] S. K. Gupta, Z. Lu, and F. Zhao. Automatic pronunciation scoring for language learning, May 15 2007. US Patent 7,219,059.
- [53] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and M. K. Sönmez. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In *INTERSPEECH*, pages 187–190, 2000.
- [54] K. Imoto, Y. Tsubota, T. Kawahara, and M. Dantsuji. Modeling and automatic detection of english sentence stress for computer-assisted english prosody learning system. *Acoustical science and technology*, 24(3):159–160, 2003.
- [55] F. Nolan. Intonational equivalence: an experimental evaluation of pitch scales. In *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona*, volume 39, 2003.
- [56] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, 2015.
- [57] N. H. De Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390, 2009.
- [58] E. Keogh. Exact indexing of dynamic time warping. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 406–417. VLDB Endowment, 2002.
- [59] M. L. Cameirao and S. G. Vicente. Age-of-acquisition norms for a set of 1,749 portuguese words. *Behavior research methods*, 42(2):474–480, 2010.
- [60] J. Greczek, E. Kaszubski, A. Atrash, and M. Matarić. Graded cueing feedback in robot-mediated imitation practice for children with autism spectrum disorders. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 561–566. IEEE, 2014.

Appendix A

Database

The present appendix shows our complete database, discriminating between training (Tr), and test (Te) subsets, and between good (green) or bad (red) labels.

Table A.1: Complete database.

Stimuli	C_1	M_1	M_2	M_3	F_1	F_2	F_3	F_4	F_5	F_6	Intonations
Banana	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Affirmation
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Question
	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Pleasure
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Displeasure
Bolo	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Affirmation
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Question
	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Pleasure
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Displeasure
Gelado	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Affirmation
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Question
	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Pleasure
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Displeasure
Leite	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Affirmation
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Question
	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Pleasure
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Displeasure
Ovo	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Affirmation
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Question
	Tr	Te	Tr	Te	Te	Te	Tr	Tr	Tr	Te	Pleasure
	Te	Tr	Te	Tr	Tr	Tr	Te	Te	Te	Tr	Displeasure

Appendix B

Recorded Utterances and Stimuli

In this appendix is present a list of all the recorded utterances in Portuguese, from all the prompts to the stimuli.

- **Generals**

"Se quiseres voltar a ouvir a explicação, clica no botão de som."

"Se quiseres voltar ao menu inicial, clica no botão casa."

"Ouve o som, clicando no botão das notas musicais."

- **Main Menu**

"Escolhe o exercício que queres fazer."

- **Exercise 1**

"As duas imagens são iguais?"

"Clica na imagem diferente."

- **Exercise 2**

"Os dois sons são iguais"

- **Exercise 3**

"Imita o que é dito."

- **Exercise 4**

"Ouve bem o alimento e escolhe se a pessoa gosta ou não"

- **Exercise 5**

"Clica na seta para cima para ouvires o exemplo de um som alto."

"Clica na seta para baixo para ouvires o exemplo de um som baixo."

"Ouve o som. É alto ou baixo"

- **Positive Reinforcement**

"Acertaste"

"Boa"

"Conseguiste"
 "É isso mesmo"
 "É mesmo assim"
 "Está quase"
 "Estás a conseguir"
 "Estás a ir bem"
 "Fantástico"
 "Já está"
 "Muito bem"

- **Negative Reinforcement**

"Acho que não está bem"
 "Ainda não está bem"
 "Não é bem assim"
 "Não está certo"
 "Oh não"
 "Oh...oh..."
 "Tenta outra vez"
 "Ups"
 "Ups não é bem assim"

- **Stimuli**

Table B.1: Recorded Stimuli.

Stimuli	Intonation 1	Intonation 2
Morangos	Pleasure	-
Bróculos	Displeasure	-
Limão	Displeasure	Affirmation
Batata	Pleasure	-
Bolacha	Pleasure	-
Couve	Displeasure	-
Papa	Pleasure	Question
Manteiga	Question	-
Mel	Displeasure	Pleasure
Massa	Question	Affirmation
Pão	Affirmation	-