Prosodic Exercises for Children with ASD via Virtual Therapy

Mariana Sofia da Silva Sousa
marianassousa@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

June 2017

Abstract

Autism Spectrum Disorders (ASD), 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 for intonation assessment 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, 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: ASD, developmental disabilities, prosodic parameters, intonation assessment, mobile application

1. Introduction

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 and it is thought that this increase in ASD diagnosis is due to a combination of two main factors: a broader definition of ASD and better efforts in diagnosis [2]. 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 [3]. 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)-IVs definition [4]. 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 [5]. 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 [6–8]. Promising methods have shown that the child's natural interests in technology can encourage communication with the therapist [9].

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.

2. Background

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 [4]. Three decades later, Kanner studied the behaviour of group of 11 children, who had in common specific clinical characteristics, never documented before. Despite the peculiarities of each individual case, Kanner was able to identify a set of common characteristics, namely: deficits in a daily basis social interactions; bizarre behaviour, characterized by restricted, repetitive and strange interests and activities; peculiar language, some children did not talk at all, and the others did pronouns exchanges or literal interpretation of the verbal information, being difficult to hold a conversation; intense and disproportionate fears to everyday noises, such as the hoover or food mixer noise [10]. Later, 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 behaviours. 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 [11]. Nowadays, the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) places both Kanner's and Asperger's definition of autism under the diagnostic of "Autism Spectrum Disorder".

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 other's moods, and anticipating other's responses [12]. 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 ADS involve impairments in reciprocal social interactions and social behaviours, several studies have investigated emotion recognition.

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. Some of these therapies include Applied Behaviour 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 [13]. When a behaviour is followed by some sort of reward, the behaviour is more likely to be repeated.

Despite not being therapies, there are two interventions/methods that are important to refer, Profiling Elements of Prosody in Speech - Communication (PEPS-C) and Picture Exchange Communication System (PECS). PEPS-C is the most used tool for evaluation of prosodic skills of children diagnosed with ASD [14]. It 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. Nowadays, the most commonly used method, while developing software for children with ASD, is the traditional PECS. PECS is an augmentative communication system, developed to help subjects in quickly acquiring a functional means of communication [15].

3. Related Work

In this work, we reviewed works concerning two distinct 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

Most of the approaches and methodologies studied have as main objective the self development, and do not concern about children communication with each other, or in adapting the tool to the user needs. The main findings provided by the surveys reviewed, 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.

As for development of vocabulary skills, 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 [15, 16]. 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 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.

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. Relating with this field, the commercial tools available nowadays help children to understand the differences of intonation 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.

Despite not being directly related with technology for children with ASD, it is important to refer a study made by Thorson et al. (2016) [17], 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.

Finalizing this section about technology for autistic children, we need to talk about VITHEA - KIDS, a platform designed for children with ASD, to develop language and generalized skills, in response to the lack of applications tailored for the unique abilities, symptoms, and challenges of autistic children.

3.2. Intonation Assessment

In this section we describe some works that helped us develop the intonation assessment method. This 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 likes and dislikes or how to express a question or an exclamation.

The state of the art in terms of intonation validation for autistic children is unfortunately very scarce. This was the main motivation for studying intonation validation in different contexts, namely in second language learning systems. This type of computer assisted language learning (CALL) systems has two large fields of research in terms of spoken language: pronunciation evaluation [18–21] and nativeness, fluency and intonation evaluation [22, 23]. Even proficient second-language speakers often have difficulty producing native-like intonation. Most of the approaches for teaching /assessing intonation take into account acoustic-prosodic 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 f0. 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.

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. The architecture of the proposed model is shown in figure 1. In the following subsections we will describe each the steps, and then present the final results for this module.

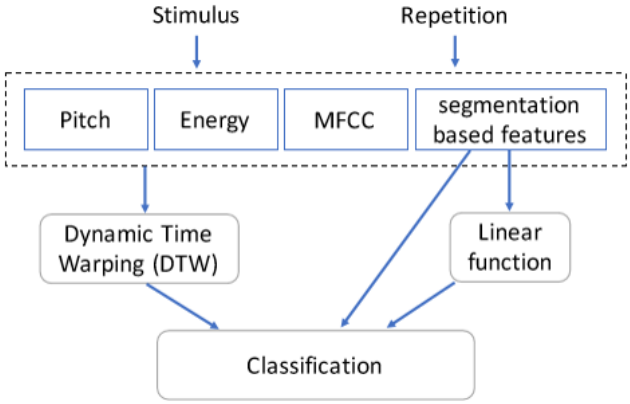

Figure 1: Intonation Assessment Method.

4.1. 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 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 exis-

tence 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.

Table 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

4.2. Feature Extraction

In accordance with several studies on automatic intonation recognition we extracted different types of prosodic features. The fundamental frequency (pitch) was computed using *Aubio*, a library to label music and sounds available as a free software. The energy contour of the speech signal was computed using a Python script. In addition, we extracted spectral characteristics in 12 sub-bands derived from the MFCCs using Librosa, which is a Python package for music and audio analysis. Finally, a set of temporal characteristics was derived by a pseudo-syllable features extraction script.

4.3. 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 [24], suppose we have two time series Q and C, of length n and m respectively, where:

$$Q = q_1, q_2, ..., q_i, ..., q_n \tag{1}$$

$$C = c_1, c_2, ..., c_i, ..., c_m \tag{2}$$

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 qi and cj (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:

$$W = w_1, w_2, ..., w_k, ..., w_K, max(m,n) \leq K <_{m+n} -1 \tag{3}$$

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 $aa' \leq 0$ and $b\text{-}b' \leq 0$. 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 $aa' \geq 0$ and $b\text{-}b' \geq 0$. 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:

$$DTW(Q,C) = min \left\{ \quad \sqrt{\sum_{k=1}^{K} w_k} \right. \tag{4}$$

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, and a cost that allow us to compute the threshold for which the intonation imitation is correct.

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

4.4. 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. 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.

For setting a threshold, we calculated the mean between both good and bad costs. 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).

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

Table 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%

$$D = \sum_{i=1}^{N} a \cdot |x_i - y_i| \quad (5)$$

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.

4.5. 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 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 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 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. The best accuracy obtained was 71.4%, with the distance function $| x^2 - y^2 |$, similarly to what happened when using pitch.

Table 3: Results obtained with Pitch.

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

Table 4: Results obtained with Power.

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

In table 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 in table 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.

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

Table 5: Results obtained with Pseudo-syllables features.

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

Table 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%**

fusion.

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.

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 some research and 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 exercises.
- The APP should score each correct answer.
- The user should be able to choose between different types of exercises.
- The user should be able to change the exercise type.
- The user should be able to finish the exercise any time.
- The APP should have reinforcements.

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 [25]. 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, a list for the affection recognition task, and another one for the imitation task. While organizing the lists we had into consideration not repeating the same intonation more than three times in a row.

5.3. Recorded Prompts

The prompt system is a set of cues with the aim of helping the player in a certain task. In [26] it is proved that graded cueing has good results and is well suited for most children with ASD. For this work we recorded positive reinforcement prompts, where the agent encourages the user to continue with the good work, and negative reinforcement prompts, where the agent encourages the user to continue the game, despise the answer being wrong. Besides, for each exercise, we recorded some specific prompts, like the exercise explanation.

5.4. New Exercises

In the main menu of our APP 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 exercises was based on a set of studied surveys. 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 witch 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.

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, 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. For the second version we display three images in the screen and the child should click in the different image.

5.4.1 Intonation Distinction

The second exercise is about intonation distinction of words. The objective of this game 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. For this task in particular, the different intonations are affirmation/question and pleasure/displeasure.

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. In order to make this task as attractive as possible, we design a cute kitty, that moves while speaking.

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. With this exercise the intention is to evaluate and develop the receptive component of the affection task. 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.

5.4.4 Up/Down Recognition

The conception of the present game was inspired in a study made by Thorson et al. (2016) [17], mentioned in 3. For our APP we made some adaptations, always focus on developing the capacity of the children with ASD of distinguish low and high sounds. There are two different versions for the present exercise. The first version consists of listening to 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. 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). In order for children to better understand this exercise, we will give, at the beginning, an example of a high and a low sound.

6. Conclusions and Future Work

As stated in the Introduction section, 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. 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.

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 a 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.

References

[1] *Diagnostic and Statistical manual of mental disorders: DSM-5*. American Psychiatric Association, American Psychiatric Association Arlington, VA, 5^{th} edition, 2013.

[2] Centers for Disease Control and Prevention. Autism spectrum disorder - data and statistics, 2014. [Online; accessed 30-December-2015].

[3] Mayada Elsabbagh, Gauri Divan, Yun-Joo Koh, Young Shin Kim, Shuaib Kauchali, Carlos Marcín, Cecilia Montiel-Nava, Vikram Patel, Cristiane S Paula, Chongying Wang, et al. Global prevalence of autism and other pervasive developmental disorders. *Autism Research*, 5(3):160–179, 2012.

[4] Guiomar Oliveira. *Edipemiologia do autismo em Portugal*. PhD thesis, Faculdade de Medicina da Universidade de Coimbra, 2005.

[5] Syamimi Shamsuddin, Hanafiah Yussof, Luthffi Ismail, Fazah Akhtar Hanapiah, Salina Mohamed, Hanizah Ali Piah, and Nur Ismarrubie 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.

[6] Brian Scassellati, Henny Admoni, and Maja Mataric. Robots for use in autism research. *Annual review of biomedical engineering*, 14:275–294, 2012.

[7] Nicole Giullian, Daniel Ricks, Alan Atherton, Mark Colton, Michael Goodrich, and Bonnie Brinton. Detailed requirements for robots in autism therapy. In *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*, pages 2595–2602. IEEE, 2010.

[8] Audrey Duquette, François Michaud, and Henri 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.

[9] Ben Robins, Kerstin Dautenhahn, R Te Boekhorst, and Aude 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.

[10] Leo Kanner. Autistic disturbances of affective contact. In *Acta paedopsychiatrica*, volume 35, pages 100–136. 1943.

[11] Hans Asperger. Autistic psychopathy in childhood. In *Autism and Asperger Syndrome*. Cambridge University Press.

[12] Thomas Owley, Laura Walton, Jeff Salt, Stephen J Guter, Marrea Winnega, Bennett L Leventhal, and Edwin 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.

[13] Joel E. Ringdahl, Todd Kopelman, and Terry S. Falcomata. Applied behavior analysis and its application to autism and autism related disorders. In Johnny L. Matson, editor, *Applied Behavior Analysis for Children with Autism Spectrum Disorders*, chapter 2, pages 15–32. Springer, 2009.

[14] Joanne McCann and Sue Peppé. Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders*, 38(4):325–350, 2003.

[15] Gianluca De Leo and Gondy 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.

[16] James 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.

[17] Jill Thorson, Steven Meyer, Daniela Plesa-Skwerer, Rupal Patel, and Helen Tager-Flusberg. Assessing prosody in minimally to nonverbal children with autism. *Speech Prosody 2016*, pages 1206–1210, 2016.

[18] Silke Witt and Steve Young. Computer-assisted pronunciation teaching based on automatic speech recognition. *Language Teaching and Language Technology Groningen, The Netherlands*, 1997.

[19] Horacio Franco, Leonardo Neumeyer, María Ramos, and Harry Bratt. Automatic detection of phone-level mispronunciation for language learning. In *EUROSPEECH*, 1999.

[20] Horacio Franco, Victor Abrash, Kristin Precoda, Harry Bratt, Ramana Rao, John Butzberger, Romain Rossier, and Federico Cesari. The sri eduspeaktm system: Recognition and pronunciation scoring for language learning. *Proceedings of InSTILL 2000*, pages 123–128, 2000.

[21] Sunil K Gupta, Ziyi Lu, and Fengguang Zhao. Automatic pronunciation scoring for language learning, May 15 2007. US Patent 7,219,059.

[22] Carlos Teixeira, Horacio Franco, Elizabeth Shriberg, Kristin Precoda, and M Kemal Sönmez. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In *INTERSPEECH*, pages 187–190, 2000.

[23] Kazunori Imoto, Yasushi Tsubota, Tatsuya Kawahara, and Masatake 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.

[24] Eamonn 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.

[25] Manuela L Cameirao and Selene G Vicente. Age-of-acquisition norms for a set of 1,749 portuguese words. *Behavior research methods*, 42(2):474–480, 2010.

[26] Jillian Greczek, Edward Kaszubski, Amin Atrash, and Maja 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.