Prosodic exercises for children with ASD via virtual therapy

Mariana Sousa[1,2], Isabel Trancoso[1,2], Helena Moniz[2,3], Fernando Batista[2,4]

[1] Instituto Superior Técnico, Universidade de Lisboa

[2] Laboratório de Sistemas de Língua Falada (L2F), INESC-ID Lisboa

[3] Faculdade de Letras/Centro de Linguística, Universidade de Lisboa

[4] Instituto Universitário de Lisboa (ISCTE-IUL)

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 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 (pitch, energy, Mel-Frequency Cepstral features, 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: ASD, Developmental Disabilities, Prosodic Parameters, Intonation Assessment, Mobile Application.

INTRODUCTION

Prosodic exercises for children with ASD via virtual therapy

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, behavioral and communication, hindering their integration into society and their relationships with others (VV., 2013). The most recent worldwide estimations, made in 2012, point to a proportion of 17 in 10,000 children with autism and 62 in 10,000 with other pervasive developmental disorders in the autism spectrum (Elsabbagh, et al., 2012). In spite of the fact that there are no recent statistics for

Portugal, a study performed in 2005 estimates that the prevalence of children diagnosed with ASD, between 7 and 9 years old, is approximately 9 in 1,000 children for Continental Portugal, according to Diagnostic and Statistical Manual of Mental Disorders (DSM) IVs definition (Oliveira, 2005).

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 (Owley, et al., 2005). Understanding emotions is a key element in social interactions, since it enables individuals to accurately recognize intentions of others and fosters appropriate responses.

The goal of this work is to develop exercises that help ASD children to understand and reproduce the emotional aspects that are reflected in communicative interaction. The original intention of this work was to integrate these prosodic exercises in VITHEA-KIDS (Mendonça, Vânia, Coheur, & Sardinha, 2015), a flexible platform that allows the easy customization of exercises for children diagnosed with ASD. Despite this original intention, the exercises were developed as a stand-alone mobile application, although it was designed for easy portability to other platforms.

This paper starts with some background on ASD therapy and related technologies. The main part of the paper describes our intonation assessment method, and corresponding results. The last sections introduce our mobile application and present some conclusions and future work.

BACKGROUND

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 Behavior Analysis (ABA), Floortime, Son-Rise, Relationship Development Intervention (RDI), among others. One of the most used therapies is ABA, which relies on the principles that explain how learning takes place, such as positive reinforcement (Ringdahl, Kopelman, & Falcomata, 2009).

In what concerns prosodic skills in ASD children, the most common assessment tool is PEPS-C (Profiling Elements of Prosody in Speech - Communication) (McCann & Peppé, 2003). This test assesses both receptive and expressive prosodic abilities. The 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. The test was adapted to European Portuguese (EP) (Filipe, 2014). In order to meet the EP characteristics, several modifications were proposed, mainly on the auditory stimulus used.

Technology for Children with ASD

Evidence is growing that technology is engaging to many children across the autism spectrum and have been shown to elicit behaviors that may not be seen in child-

person interactions (Scassellati, Admoni, & Matarić, 2012) (Giullian, et al., 2010) (Duquette, Michaud, & Mercier, 2008).

The majority of the reviewed technological work for children with ASD is inspired in Picture Exchange Communication System (PECS) (De Leo & Leroy, 2008), which is simple to use and even possible to customize, although the possibilities are quite limited. Our survey of the state of the art showed that most studies target vocabulary enrichment (Moore & Calvert, 2000) (Bosseler & Massaro, 2006), and communication skills in a social context. (Mitchell, Parsons, & Leonard, 2007) (Parsons, Mitchell, & Leonard, 2004) (Cihak, Smith, Cornett, & Coleman, 2012) (Ohene-Djan, 2010). Most approaches adopt visual and auditory stimuli to motivate children. Some of them explore the use of virtual environments for increased independent communicative interactions. Other approaches allow children to choose and combine several images to form a message (De Leo & Leroy, 2008).

The commercial tools (such as Learning with Rufus - Feelings and Emotions; Emotions and Feelings – Autism; SPEAKall!; iCommunicate and TalkInPictures) 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.

VITHEA – KIDS (Mendonça, Vânia, Coheur, & Sardinha, 2015) is a platform specifically 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. The types of exercises present in VITHEA – KIDS are: 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, which could be a picture, a text or a video, and a set of possible answers, in which only one of the answers is correct. The platform allows caregivers to build customized multiple choice exercises while taking into account specific needs/characteristics of each child. The set of exercises does not yet include exercises for improving the intonation skills of autistic children, which gave us additional motivation for this work.

INTONATION ASSESSMENT METHOD

Prosodic exercises can either help children distinguish between different intonations or mimic certain intonations. This section concerns the latter type of exercises, which are the most challenging ones to implement. The goal of our intonation assessment method is to evaluate and develop the child skills to imitate different intonations, for short stimuli (words).

Related Work in Intonation Assessment

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 (Franco, Neumeyer, Ramos, & Bratt, 1999) (Franco, et al., 2000) (Gupta, Lu, & Zhao, 2007) and nativeness, fluency and intonation evaluation (Teixeira, Franco, Shriberg, Precoda, & Sõnmez, 2000) (Imoto,

Tsubota, Kawahara, & Dantsuji, 2003). 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, energy and Mel Frequency Cepstral Coefficients (MFCCs), 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, which can even bypass the feature extraction stage, in end-to-end classification tasks. However, this type of approach requires very large training databases, and this is one of the major limitations in our work.

DATA COLLECTION

The first step of our data collection was to record an adult European Portuguese (EP) female speaker uttering a total of 20 stimuli (shown in Table 1), consisting of isolated words with different intonations. Ideally, the second step should involve recording imitations of the different stimuli by both autistic and healthy children, and the corresponding labels should be done by therapists. At this preliminary stage, however, we tried to validate the method using only healthy subjects. We recorded 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. 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. For each participant, the set of 20 utterances was randomly subdivided into two subsets, one for training our algorithm, and another one for testing it. Table 2 shows the complete dataset, discriminating between training and test subsets, and between good or bad labels.

Table 1. Database stimuli

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

Note: This table presents the recorded stimuli, as well as the intonations for each stimulus word, that composes our database.

Table 2. 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

Stimuli	C_1	M_1	M_2	M_3	F_1	F_2	F_3	F_4	F_5	F_6	Intonations
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

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

ARCHITECTURE

The architecture of the proposed module is shown in Figure 1. The following subsections describe each one of the blocks in more detail.

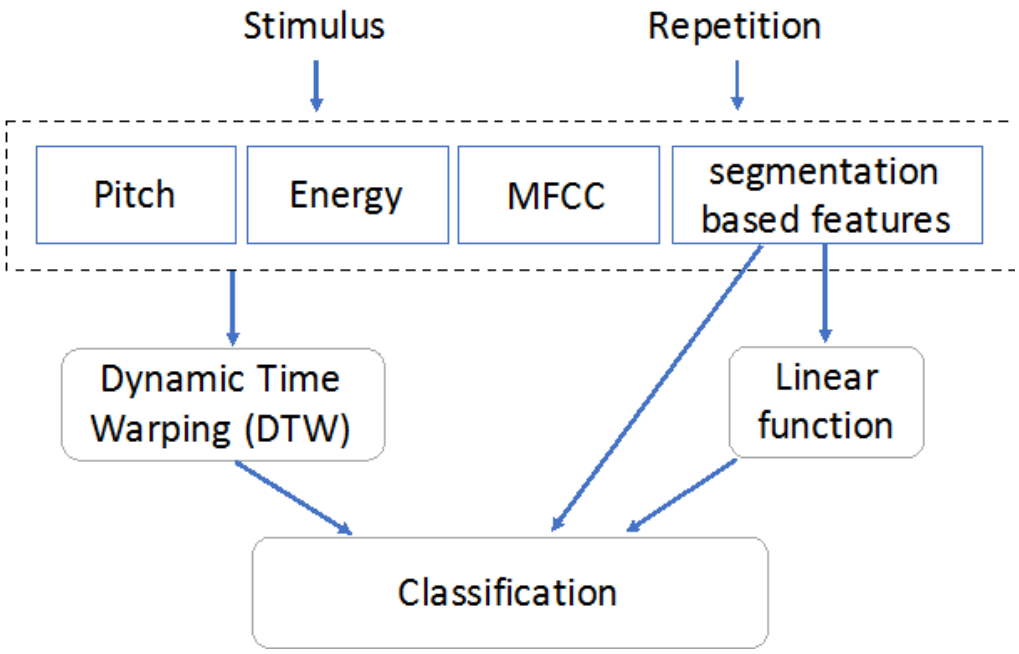

Figure 1. Intonation Assessment Method Architecture

Feature Extraction

In accordance with several studies on automatic intonation recognition, different types of prosodic features were extracted. The fundamental frequency (pitch) was computed using Aubio (Brossier, 2006), a library to label music and sounds available as a free software. The energy of the speech signal was computed using the snack sound toolkit (http://www.speech.kth.se/snack/), by means of the WaveSurfer software (Sjölander & Beskow, 2000). In addition, we extracted spectral characteristics in 12 sub-bands derived from MFCCs using Librosa (McFee, et al., 2015), which is a Python package for music and audio analysis. Finally, a set of segmentation- based characteristics was derived by means of freely available a pseudo-syllable features extraction script (De Jong & Wempe, 2009). It includes the number of pseudo-syllables, number of pauses, speech duration, phonation time, speech rate, and articulation rate.

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. Our DTW module was an existent Python module that yields the optimal warping path and a cost, and allows using a user-defined cost function.

Classification

When performing experiments with one feature only, our classification is based on a threshold, previously calculated used the existing training set. We start by calculating the distance between all the repetitions in the training set and their corresponding stimulus. Based on the achieved values, we compute the mean and standard deviation for the two different targets (Good and Bad). The threshold is then defined as the mean between the two centroids, taking into account the standard deviation of their values.

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.

RESULTS

In this section, we present the results of evaluating the developed method, applied separately to each set of features, in order to know what feature is more informative in terms of intonation imitation. Once the threshold was tuned with the data of the training set, each utterance of the test set was classified as good or bad imitation, obtaining the correspondent "correct" (C) or "incorrect" (I) labels for each utterance. The accuracy of the algorithm for each feature set was computed as the percentage of correct classifications.

The results of applying the algorithm separately for each extracted feature an also for the fusion of all features are presented in Table 3. 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.

Table 3. Accuracy results of the intonation assessment method

Feature		**Accuracy**	
		Mean&stdev classifier	Decision Tree
Framed-based DTW	MFCCs	**83.3%**	82.2%
	Pitch	72.2%	72.2%
	Energy	70%	74.4%
	Fusion	--	**77.8%**
Segment-based	Pseudo-syllable features	--	73.3%
	Fusion	--	75.5%

Note: This table comprises the final results of the implemented algorithm for the intonation assessment. The accuracy of the algorithm was separately evaluated for each set of features, as displayed in the table.

NEW CONTRIBUTIONS TO THE VIRTUAL THERAPISTS

For our application we propose four exercises for prosody training of children diagnosed with autism, in order 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. These exercises will be described in the order they should be followed.

Intonation Distinction

This exercise is about intonation distinction of words/sentences. The objective of this exercise is to evaluate and develop the skill of understanding intonation changes in short stimulus (words) and long stimulus (sentences). For this task, the discrimination paradigm of "equal vs different" is used and the procedure consists of 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.

UP/DOWN RECOGNITION

This exercise focus on developing the capacity of the children with ASD of distinguishing low from high tones. We developed two versions of this exercise. The first version consists of listening to a single sound and then pressing the up arrow for "high frequency" sounds or the down arrow for "low frequency" sounds. The second exercise is a little more complex, since a sequence of two sounds is played 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 the exercise, we give, at the beginning, an example of a high sound and a low sound.

AFFECT RECOGNITION

This exercise is concerned with the understanding and use of prosody to express pleasure or displeasure. An image of a food item appears on the screen, followed by an auditory stimulus, which is the food item name pronounced with pleasure or displeasure. The answer consists of select one of the two buttons that appear on the screen simultaneously, one with a smiley face (pleasure) and another one with a sad face (displeasure).

INTONATION IMITATION

The intonation imitation exercise objective is to develop the children skills to imitate different types of intonations in short stimuli, composed by one word. This exercise integrates the above mentioned intonation assessment method, and its purpose is to allow children to have more confidence when expressing themselves with emotion, or to express their tastes while interacting with someone.

CONCLUSIONS

In this work we presented a set of prosodic exercises in virtual therapists for children with ASD, as a complement of therapy sessions, with the aim of developing and enriching the prosodic abilities of such audience, thus contributing to a better communicative interaction. In order to achieve our goals, we proposed an intonation

assessment method, with the objective of classifying the intonation imitation produced by autistic children. 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. The paper concludes by a brief description of the set of prosodic exercises that we have implemented in our mobile application.

This work should be pursued in several directions, the first being the evaluation of the potential improvements of classifier fusion for intonation validation. Tuning the classifier with a database of autistic children is a must, as well as integrating the new exercises in VITHEA- KIDS, and finally evaluating user satisfaction with this community.

Author note

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and Post-doc grant SFRH/PBD/95849/2013. This work is also supported by Project RAGE, European Union Horizon 2020 Programme for Research and Innovation under grant agreement 644187, and by Quality Assurance Project between INESC-ID and Unbabel.

REFERENCES

Bosseler, A., & Massaro, D. (2006). 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.

Brossier, P. (2006). *The aubio library at mirex 2006*. MIREX.

Cihak, D., Smith, C., Cornett, A., & Coleman, M. (2012). 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.

De Jong, N., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385-390.

De Leo, G., & Leroy, G. (2008). Smartphones to facilitate communication and improve social skills of children with severe autism spectrum disorder: special education teachers as proxies. *Proceedings of the 7th international conference on Interaction design and children*, (pp. 45-48).

Duquette, A., Michaud, F., & Mercier, H. (2008). Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Autonomous Robots*, 24(2), 147-157.

Elsabbagh, M., Divan, G., Koh, Y. ‐ J., Kim, Y., Kauchali, S., Marcín, C., . . . Wang, C. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism Research*, 5(3), 160-179.

Filipe, M. (2014). Prosodic Abilities in Typpically Developing Children and those Diagnosed with Autism Spectrum Disorders - Clinical Implications for Assessment

and Interventions. University Porto - Faculdade de Psicologia e de Ciências da Educação.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., . . . Cesari, F. (2000). The SRI EduSpeakTM system: Recognition and pronunciation scoring for language learning. *Proceedings of InSTILL 2000*, 123-128.

Franco, H., Neumeyer, L., Ramos, M., & Bratt, H. (1999). Automatic detection of phone-level mispronounciation for language learning. EUROSPEECH.

Giullian, N., Ricks, D., Atherton, A., Colton, M., Goodrich, M., & Brinton, B. (2010). Detailed requirements for robots in autism therapy. *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on* (pp. 2595-2602). IEEE.

Gupta, S., Lu, Z., & Zhao, F. (2007). Automatic pronunciation scoring for language learning. Google Patents.

Imoto, K., Tsubota, Y., Kawahara, T., & Dantsuji, M. (2003). Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system. *Acoustical Science and Technology*, 24(3), 150-160.

McCann, J., & Peppé, S. (2003). Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders*, 38(4), 325-350.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. *Proceedings of the 14th Python in Science Conference*.

Mendonça, Vânia, Coheur, L., & Sardinha, A. (2015). Vithea-kids: a platform for improving language skills of children with autism spectrum disorder. Proceedings of the *17th International ACM SIGACCESS Conference on Computers & Accessibility*, (pp. 345- 346).

Mitchell, P., Parsons, S., & Leonard, A. (2007). Using virtual environments for teaching social understanding to 6 adolescents with autistic spectrum disorders. *Journal of Autism and Developmental Disorders*, 37(3), 589-600.

Moore, M., & Calvert, S. (2000). Brief report: Vocabulary acquisition for children with autism: Teacher or computer instruction. *Journal of Autism and Developmental Disorders*, 30(4), 359-362.

Ohene-Djan, J. (2010). Winkball for schools: An advanced video modelling technology for learning visual and oral communication skills. In *Advanced Learning Technologies (ICALT)* (pp. 687-689). IEEE.

Oliveira, G. G. (2005). Epidemiologia do autismo em Portugal: um estudo de prevalência da perturbação do espectro do autismo e de caracterização de uma amostra populacional de idade escolar. **Doctoral thesis**.

Owley, T., Walton, L., Salt, J., Guter, S. J., Winnega, M., Leventhal, B. L., & Cook, E. H. (2005). An open-label trial of escitalopram in pervasive developmental disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44(4), 343-348.

Parsons, S., Mitchell, P., & Leonard, A. (2004). The use and understanding of virtual environments by adolescents with autistic spectrum disorders. *Journal of Autism and Developmental Disorders*, 34(4), 449-466.

Ringdahl, J. E., Kopelman, T., & Falcomata, T. S. (2009). Applied behavior analysis and its application to autism and autism related disorders. In J. Matson, *Applied behavior analysis for children with autism spectrum disorders* (pp. 15-32). Springer.

Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14, 275-294.

Sjölander, K., & Beskow, J. (2000). Wavesurfer-an open source speech tool. *INTERSPEECH*, (pp. 464-467).

Teixeira, C., Franco, H., Shriberg, E., Precoda, K., & Sõnmez, K. (2000). Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. *INTERSPEECH*, (pp. 187-190).

VV., A. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®).* American Psychiatric Pub.