Proceedings of the RaPID-4 @LREC 2022, pages 22–30
Marseille, 25 June 2022

Extraction and Classification of Acoustic Features from Italian Speaking Children with Autism Spectrum Disorders

Federica Beccaria[1], Gloria Gagliardi[1], Dimitrios Kokkinakis[2]
[1] University of Bologna – Department of Classical Philology and Italian Studies
[2] University of Gothenburg – Department of Swedish, Multilingualism, Language and Technology
federica.beccaria@studio.unibo.it, gloria.gagliardi@unibo.it, dimitrios.kokkinakis@svenska.gu.se

Abstract

Autism Spectrum Disorders (ASD) are a group of complex developmental conditions whose effects and severity show high intraindividual variability. However, one of the main symptoms shared along the spectrum is social interaction impairments that can be explored through acoustic analysis of speech production. In this paper, we compare 14 Italian-speaking children with ASD and 14 typically developing peers. Accordingly, we extracted and selected the acoustic features related to prosody, quality of voice, loudness, and spectral distribution using the parameter set eGeMAPS provided by the openSMILE feature extraction toolkit. We implemented four supervised machine learning methods to evaluate the extraction performances. Our findings show that Decision Trees (DTs) and Support Vector Machines (SVMs) are the best-performing methods. The overall DT models reach a 100% recall on all the trials, meaning they correctly recognise autistic features. However, half of its models overfit, while SVMs are more consistent. One of the results of the work is the creation of a speech pipeline to extract Italian speech biomarkers typical of ASD by comparing our results with studies based on other languages. A better understanding of this topic can support clinicians in diagnosing the disorder.

Keywords: Autism Spectrum Disorders, acoustic analysis, machine learning, openSMILE, eGeMAPS

1. Introduction

The American Psychiatry Association defines Autism Spectrum Disorders (ASD) as a group of complex developmental conditions whose effects and severity are different in each person. However, some common symptoms have been found whose presence represents the criteria used during the diagnosis. According to the DSM-5, one of them is the presence of impairments in social communication (Criterion A). Thus, the quality of language is an essential indicator during the diagnosis of ASD, both in comprehension and production. Even if these linguistic characteristics are present in a spectrum that showcases a wide variety, they still have something in common. Social interaction is mainly completed using different language skills.

In the present study, we investigated the speech production of Italian-speaking children with ASD to understand if there are acoustic features that can be shared along the spectrum. Indeed, there are already many English contributions showing abnormalities of autistic speech at the prosodic level. Unfortunately, there are few studies on this promising field in Italian. To conduct this investigation, we performed an acoustic feature extraction and a supervised learning classification between the speech production of Italian-speaking children with ASD and their peers with typical neurodevelopment (TD).

2. Prosody in ASD Speech

The study of prosodic traits in people with autism is relatively new and, compared to other linguistics domains, still little explored (Diehl & Paul, 2013; Kiss et al., 2012; Tanaka et al., 2014; Van Santen et al., 2010). As a result, this research field was called by some experts the “Cinderella of speech” that remains "in the cellar, with few visitors" (Crystal, 2009; p. 257). Nevertheless, the research on the typical acoustic features in people with different neurodevelopmental disorders is promising. Through various methods, usually based on multimodal investigations (i.e., behavioural assessments, acoustic analysis, electrophysiological measures, brain imaging), it has been demonstrated that the speech of autistic people shows some anomalies from the prosodic point of view. Indeed, common variabilities along the spectrum have been recorded in the movements and the pitch types produced (Shriberg et al., 2001). This acoustic pattern is the speakers' attitude and emotional status medium.

Based on these features, two main prosodic behaviours are commonly identified during the speech act: the pragmatic (or linguistic) and the affective functions (Anolli, 2002). The first represents the illocutive force, which is the act itself of talking by the speaker (see Searle & Vanderveken, 1985). Moreover, it distinguishes the type of sentence produced, e.g., interrogative or affirmative. On the other hand, the second function represents the medium - sometimes unintended - of the emotional status felt by the speaker. Thus, people with alterations of these prosodic productions may exhibit impairments in elaborating the vocal chants and sentences showing their emotional status. Moreover, these impairments affect their comprehension of other people, causing difficulties in social interaction and communication (Olivati et al., 2017).

From its first descriptions, the speech of people with autism has been defined as being monotonous, robotic, and pedant (Kanner, 1943). The patients present difficulties both in the production and perception skills. For instance, Kanner (1943; p. 228) wrote about one of the children he studied: "It made no difference whether one spoke to him in a friendly or a harsh way". Thus, the scholar who first defined autism gave an implicit focus on prosodic and affective traits in the speech production and comprehension of people with the disorder. However, the researchers ignored this part of Kanner's study in the decades that followed. Nevertheless, through prosody, we can detect the acoustic patterns that show the speaker's emotional status, one of the most visible symptoms in the atypical communication of people with ASD. Thus, during the last years, the research moved to the study of acoustic correlates while analysing these typical features of the disorder.

3. The Dataset

The participants in the present study come from a pool of Italian-speaking children in a homogeneous geographical area from the region between Florence, Pistoia, and Prato. The corpus consists of audio recordings collected from two cohorts: children with ASD and their peers with typical neurodevelopment (TD). The data are balanced on the number of participants and their demographic characteristics. The children are 14 for each group with the same age (from 6 to 10 years) and sex (11 M, 3 F). Gender disparity is taken from the epidemiology of the disorder recorded by the DSM-5 (APA, 2013), i.e., four males every one female.

The participants in the study were recruited from a previous project on discourse and storytelling in autism (Biancalani, 2019), where the children were asked to tell a story from six pictures stimulating a semi-spontaneous speech during the interviews. The images illustrate a story about a birthday party and are easily interpretable by neurotypical children of different ages. The pictures come from the toy Shubi collection *Storie da raccontare* (in English, 'Story to tell'). The children from the ASD group were recruited from the speech and language therapy service of AUSL Toscana Centro and the Onlus foundation "Opera Santa Rita". The diagnosis was made by a neuropsychiatrist according to DSM-5 criteria. The data collection was carried out by a designed speech therapist in June 2019, after receiving the written consent of the caregivers of the children. The recordings were realised with a video camera placed on a tree-legged support. The setting was designed so that the child would feel comfortable. Therefore, the meeting was conducted in the room where they usually play. The interview started with activities generally done during the treatment session so the child would act in the most spontaneous way.

It was necessary to conduct new data collection for the TD group. The recording was done for qualitative analysis in the previous study, and the audio quality was not good enough to realise an acoustic investigation. In particular, the background noise was so high that it was impossible to identify the child's voice automatically. It was attempted to denoise the recordings with the software Audacity, but this solution would have significantly changed the shape of the waveforms and their quality in general. The participants were chosen from the same geographical area and had the same demographic characteristics as those from the ASD group. Moreover, due to the COVID-19 pandemic, it was impossible to collect the data *in situ*, so the parents of each child did the recording using their phones. Even though we are aware that this might eschew our results, we consider that it will not have that significant impact because the storytelling task remained the same, and the quality of the recordings was high (i.e., there was no background noise).

4. Extraction, Selection, and Classification of Acoustic Features

In the present study, we decided to use the *Munich open-Source Media Interpretation by Large feature-space Extraction* (openSMILE) to extract the acoustic features. In the area of autistic vocalisation detection, this software has been used in previous studies, reaching satisfying results (Asgari & Shafran, 2018; Cho et al., 2019; Kim et al., 2017; Lee et al., 2013; Li et al., 2019; Marchi et al., 2015; Pokorny et al., 2017). After extracting the acoustic features, we selected the most statistically significant between the two groups of our dataset (ASD and TD). Then, we tested the features selection by implementing machine learning algorithms with a binary classification task. The role of training supervised learning methods is to classify them and show if they are significant in the speech production of people with ASD, according to the performance of each model. This model may evolve, through further studies, into a tool that helps the clinician determine whether Italian children have ASD at a young age.

4.1 Methods

We used openSMILE version 2.1, developed for the Interspeech challenge (Schuller et al., 2013). Among the feature sets currently available – i.e., GeMAPS (Geneva Minimalistic Standard Parameter Set), eGeMAPS (Eyben et al., 2015), and ComParE (Schuller et al., 2016) - we applied the second that was specifically ideated by its developers to become a tool used in paralinguistics and clinical speech analysis.

Moreover, we chose this feature set over the other two proposed by openSMILE for several reasons. First, we decided against using ComParE, given that the size of the feature space (n = 6376) vastly outnumbers the sample size of our dataset. Furthermore, this would have caused our machine learning models to overfit, which is highly undesirable. On the other hand, we chose eGeMAPS over GeMAPS, given that the former extracts features based on their relation to various psychological changes in voice production (Eyben et al., 2015), which has proven useful in previous studies (Julião et al., 2020; Lee et al., 2020; Marchi et al., 2015; Memari et al., 2020; Pokorny et al., 2017; Ringeval et al., 2016; Rybner et al., 2022; Schmitt et al., 2016). The acoustic features extracted by eGeMAPS are related to the frequency, energy, amplitude, and distribution on the spectrum. These are presented in Table 1 - 3 with a short explanation extracted from Eyben et al. (2015; pp. 4-5).

Features	Explanation
Pitch	Logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
Jitter	Deviations in individual consecutive F0
Formants 1, 2, 3 frequencies and bandwidth	The centre frequency and the bandwidth of the first, second, and third formant

Table 1: Frequency related features

Features	Explanation
Harminics-to-Noise Ratio	Relation of energy in harmonic components to energy in noise-like components
Loudness	Estimate of perceived signal intensity from an auditory spectrum
Shimmer	Difference of the peak amplitudes of consecutive F0 cycles

Table 2: Energy and amplitude related features

Features	Explanation
Alpha Ratio	Ratio of the summed energy from 50–1000 Hz and 1–5 kHz
Formants 1, 2, 3 with relative energies	Ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F0
Hammarberg Index	Ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region
Harmonic difference H1–H2	Ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2)
Harmonic difference H1–A3	Ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3)
MFCC 1-4	Mel-Frequency Cepstral Coefficients 1-4
Spectral flux	Difference of the spectra of two consecutive frames
Spectral Slope 0–500Hz and 500–1500 Hz	Linear regression slope of the logarithmic power spectrum within the two given bands

Table 3: Spectral (balance) related features

Once the acoustic features presented in Tables 1 - 3 were extracted, we selected the most statistically significant ones by implementing a non-parametrical statistical test, namely Mann-Whitney U-test (Mann & Whitney, 1947; Wilcoxon, 1945). Thus, with the Mann-Whitney U-test, we tested whether the features had similar values between the two groups and selected the ones presenting the more significant distance. In the final discussion, we introduced the features selected by comparing them with those obtained by other studies. In doing so, we considered the different methods applied to the data collection and the analysis itself.

Having irrelevant features can decrease the accuracy of machine learning models, especially when dealing with linear models such as support vector machines. On the other hand, the feature selection operated on clinical recorded data can lead to high performances of the classifier, which are not reproducible on new data considering all the speech production. This aspect produces a bias in the results obtained by the classifier when used to create a tool able to distinguish the speech productions of the disease. For instance, while performing the same task on people with Alzheimer Dementia's speech, Luz et al. (2020) report that the performance evaluation metrics drop consistently if applied to the same dataset without performing the feature selection. However, this work aims to test the feature extraction effectiveness and not automatically classify ASD from new speech data. This goal could be reached by future studies conducted on larger datasets.

We pre-processed the data obtained through feature selection to prepare it for the supervised learning methods. Then we normalised the data and implemented the K-fold Stratified cross-validation to train the models (k = 3). Thus, we split the training set into k parts, denominated folds. Next, we train a model that uses that fold as a validation set and the rest as its training set for each fold. This helps avoid overfitting the noise in the data. We split 80% for the train and 20% for the test sets. Given the small number of samples on our corpus (ASD = 14, TD = 14, total features = 16), the data processed on the sets were 22 by the train and six by the test sets.

The machine learning methods implemented are all supervised: Decision Trees (DTs), K-Nearest Neighbours (KNNs), Random Forests (RFs), and Support Vector Machines (SVMs). First, we evaluate the performances of each model trained using different metrics: accuracy, recall, precision, F1-score, and Area under the Curve (AUC). Then, we chose the best ten models obtained by running each supervised method and comparing them with the others. Finally, we selected the best performing model of each method.

All the computational steps are done by implementing different algorithms in Python (Chollet, 2021; Downey et al., 2012; Van Rossum & Drake, 2011) with the aid of the Jupyter Notebook (Kluyver et al., 2016). Moreover, all the machine learning methods performed are implemented using the Scikit-learn module for the Python programming language (Pedregosa et al., 2011). The code used is publicly available on GitHub: https://github.com/federica-bcc/speech-autism.

4.2 Results

We extracted 88 parameters concerning the frequency, energy, and spectral distribution. Table 4 reports the parameters selected with their respective functionals in parenthesis (μ = mean, σ = standard deviation), the values obtained from both the groups with the number of outliers in parenthesis if found. The last column indicates the significance levels through the p-value (*p*).

The best models of each supervised method reach high accuracy, with the highest being Decision Tree, Random Forest, and Support Vector Machine (accuracy = 83%), while the lowest KNN (accuracy = 67%). The AUC metric's highest values are reached by DT and SVM (AUC = 88%), while the KNN and the RF have lower performances (AUC = 75%). Tables 5 and 6 report the results obtained by the best model of each classifier on these metrics and the others (recall, precision, and F1-score), both on the train and the test sets, respectively.

On the other hand, Table 7 reports the mean of the evaluation metrics obtained by all the models for each method giving a clearer view of their overall behaviour on the classification task.

4.3 Discussion

In the present study, we analysed the speech production of Italian speaking children with ASD. Our corpus comprises 28 audio recording files divided into two groups: 14 children with ASD and 14 controls. First, we implemented the acoustic feature extraction using eGeMAPS provided by the openSMILE toolkit. Next, we extracted 88 parameters for each audio file and selected the most statistically significant between the two groups. Finally, we implemented four supervised learning algorithms to test the validity of the feature selection.

In the following sections, we discuss the features obtained with the feature selection (Section 4.3.1) and the results from the classification task (Section 4.4.2).

Feature	ASD	TD	p-value
Pitch falling slope (σ)	168.81 ± 61.29 (0)	137.27 ± 130.12 (2)	0.0409*
F2 Frequency (σ)	0.17 ± 0.013 (3)	0.15 ± 0.017 (0)	0.0030**
F2 bandwidth (σ)	0.33 ± 0.038 (0)	0.38 ± 0.075 (1)	0.0326*
Jitter (μ)	0.036 ± 0.008	0.024 ± 0.011	0.0094**
Jitter (σ)	1.69 ± 0.21 (2)	1.88 ± 0.26 (1)	0.0094**
Shimmer (μ)	1.23 ± 0.067	1.04+ 0.14	0.0016**
Shimmer (σ)	0.46 ± 0.02 (1)	0.66 ± 0.16 (0)	0.0010***
Harmonics-to-Noise Ratio (μ)	5.04 ± 1.14	7.88 ± 2.10	0.0012**
Harmonics-to-Noise Ratio (σ)	1.03 ± 0.29	0.64 ± 0.28	0.0016**
Loudness (μ)	1.16 ± 0.44	0.82 ± 0.30	0.0508
Loudness rising slope (σ)	7.87 ± 2.25	0.04 ± 0.02	0.0409*
Loudness (Percentile 20.0)	0.59 ± 0.21	0.32 ± 0.18	0.0035**
Spectral Flux (μ)	0.72 ± 0.38	0.42 ± 0.24	0.0366*
Spectral Flux voiced segments (μ)	0.86 ± 0.43	0.51± 0.26	0.0409*
Slope unvoiced segments, 0-500 Hz (μ)	0.056 ± 0.02	0.04 ± 0.02	0.0456*
Slope unvoiced segments, 500-1500 Hz (μ)	-0.0096 ± 0.0022 (0)	-0.0027 ± 0.0071 (1)	0.0026**

Table 4: Values of the acoustic features with statistical significance between the ASD and TD groups.
Results are expressed as *means ± standard deviations (n. outliers)*. Asterisks indicate when the group-related difference is significant under the Mann-Whitney U-test: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$.

Methods	Acc.	Rec.	Prec.	F1-sc.	AUC
DT	82	92	100	96	96
KNN	77	83	77	80	77
RF	77	67	89	67	78
SVM	73	83	71	77	72

Table 5: Values of the evaluation metrics obtained on the train set by the best models

Methods	Acc.	Rec.	Prec.	F1-sc.	AUC
DT	83	100	96	80	88
KNN	67	100	50	67	75
RF	83	50	100	67	75
SVM	83	100	67	80	88

Table 6: Values of the evaluation metrics obtained on the test set by the best models

Methods	Acc.	Rec.	Prec.	F1-sc.	AUC
DT	68.2 (16.40)	100 (0)	54.5 (13.50)	69.5 (11.46)	76.3 (12.92)
KNN	58.5 (12.02)	75 (26.35)	42 (10.05)	52.1 (11.83)	59.8 (12.66)
RF	70.1 (10.33)	65 (24.15)	64 (25.03)	59.2 (8.48)	68.5 (6.85)
SVM	79.8 (6.75)	95 (15.81)	63.6 (7.17)	77.4 (5.48)	84.9 (5.44)

Table 7: Means and (sd) of the evaluation metrics obtained on the test set by running the models ten times

4.3.1 Typical Acoustic Features of ASD Speech

To discuss the features selected, we divided them into thematic groups according to their linguistic qualities: frequency-related parameters (pitch and second formant), voice quality (shimmer, jitter, and Harmonics-to-Noise Ratio), loudness, and spectrum-related parameters (spectral flux and slope).

Frequency-related parameters. The pitch is one of the main features of prosodic analysis. In general, there are opposite findings in the literature regarding the mean values of the pitch. The present study found higher values in this functional, but the difference between the two groups is not statistically significant. However, the increasing pitch could suggest chasing away the idea that the speech of people with ASD is robotic, monotonous and without melodic variation, as reported in the past literature (Kissine & Geelhand, 2019; Nayak et al., 2019; Olivati et al., 2017).

We found interesting results on the standard deviation of the pitch falling slope. Moreover, even if not statistically significant, the same pattern is observed in the rising slope. These results confirm the high variation in the general prosody production, specifically on the intonation contours. Sharda et al. (2010) related these pitch excursions to their range and showed similarities to the one observed during "motherese" speech postulating a delayed developmental trajectory of speech. Nevertheless, even if interesting, Bonneh et al. (2011) disproved these results and showed that this trend does not always hold.

Another explanation can be found in some interesting results on impairments controlling the cortical pitch. The most relevant finding on these assumptions is related to auditory processing in autism (Boddaert et al., 2004; Rosenhall et al., 1999). The research in this field has increased in the past few years, mainly thanks to neuro-imaging techniques applied to experiments through a multimodal optic study.

Moreover, the extreme variation in the pitch values shows the difficulties of people with autism to perceive and, consequently, produce prosody in the same way as their peers (Olivati et al., 2017). This can lead to many difficulties in social interaction and communication because of the lack of others' speech intention comprehension than the production itself (Bonneh et al., 2011). Moreover, on the correlation between production and perception, many studies conducted on the comprehension of the emotions communicated by the interlocutor show that children with ASD have less capacity than their peers. The same trend is reflected in their speech, especially in using different intonations to transmit each emotion (Chiew et al., 2017; Hubbard et al., 2017; Schelinski & Kriegstein., 2019).

Finally, another problem described in the literature is the influence of external factors on the recording and the inhomogeneity in extracting the correlations. First, the feature related to the formants, including the fundamental frequency, is sensitive to the speaker's age, gender, and height (Bone et al., 2014). Second, as reported in McCann and Peppe (2003), it would be expected that these descriptors for prosodic abnormalities should appear in many studies. However, the findings do not show coherent discussions because the evaluation measures are not well defined.

For these reasons, in future studies, it will be interesting to investigate both the correlation of voice features to the personal characteristics of the speakers (age, gender, and height) and compare all the results with other studies that used the same metrics.

Voice quality. The voice quality is measured as the difficulty in controlling the vocal fold vibrations, transforming the production into hoarseness, breathiness, and creaky voice. These irregularities can be quantified through the analysis of some features that "reflect mathematical properties of the sound wave" (Robin et al., 2020; p. 102), such as the jitter for the pitch, the shimmer for the intensity and the Harmonics-to-Noise Ratio (HNR) for the description of the periodic and aperiodic acoustic propagation (Tsanas et al., 2011). The present study found interesting results on all these parameters, confirming the observations drawn by other investigations.

Jitter and shimmer are related: the first measures periodicity in the speech signal, while the second the difference from a cycle to the next one. As done in other studies, these values were calculated using the local method by evaluating the pitch period and magnitude once per each span of the period (Boersma, 2001; Bone et al., 2012). However, they are also related to another aspect: they are valuable parameters to measure in a speech pathology analysis because the voice with language impairment is likely to have higher values than a healthy one (Styler, 2021).

In the present study, two populations are compared with the assumption that one of them (ASD) shows impairments in vocal production. The results we obtained from the jitter and the shimmer confirm this hypothesis. Indeed, the means are higher in the speech of ASD, with statistically significant differences described by the p-values obtained by the Mann-Whitney U-test (0.0094 for jitter and 0.0016 for shimmer). However, the results are the opposite for the standard deviations. The findings suggest that these acoustic features vary consistently less in children with autism but have higher values on average.

This trend has also been observed in previous studies. For instance, in Kissine & Geelhand (2019), the authors noted a highly statistically significant difference between these two parameters, with a higher rate in the production of ASD (jitter $p < 0.001$; shimmer $p = 0.001$). Moreover, their sample was composed by adults (mean age: about 28 years old) while, in the present study, the participants were children. Hence, future studies might explore if this trend is typical of autism throughout life, meaning a turning point in the early diagnosis of the disease. Indeed, the analysis of these correlations, combined with the pitch, intensity, and pause count, supports the hypothesis of assessing the speech modulation in ASD through studying the measure of dynamic-intonation variability (Bone et al., 2015).

Moreover, the jitter and the shimmer show the noise present in the speech, and their values can be sensitive to its presence in the recording. For this reason, it is essential to analyse also the HNR that usually detects the friction in the vocal tract, attributed to hoarse, breathy, or laryngeal pathologies when it decreases significantly (Styler, 2021).

In Bone et al. (2014), the mean of the HNR is shown to be strictly related to the jitter: when this latter increases, the other decreases. In the present study, we found this trend with significant results both on the means (jitter: p = 0.0094**, HNR: p = 0.0012**) and on the standard deviations (jitter: p = 0.0094**, HNR: p = 0.0016**) that follows the opposite growth for both the mean and the standard deviation.

The negative correlation between jitter and HNR is observed in many studies concerning the analysis of breathless, hoarseness and roughness voices, where they are also correlated to an increase in the cepstral values (Halberstam, 2004; Hillenbrand et al., 1994). McAllister et al. (1998) correlated in their studies the jitter to the breathy, hoarse, nasal speech and the shimmer to the breathiness, but no correlation with the cepstral values was found in this type of speech. In the same way, the present study did not find statistically significant results on these latter features. Bone et al. (2014) reported the same trend that we obtained. Therefore, we agree with the authors that it is necessary to conduct more analysis regarding voice quality to confirm this trend and to be able to use it during the diagnosis (Bone et al., 2014: 1173).

The loudness. The loudness is defined as the energy intensity produced by a sound wave. We found a statistically significant difference on the 20th percentile (p = 0.0035), in the standard deviation (p = 0.0409), and in the general mean (p = 0.0508). Even if these functionals measure different distribution aspects, they all present the same trend, showing higher values for the ASD group. These results are confirmed in Bone et al. (2012), where the role of intensity in the perception of abnormal volume is underlined with the increasing rate of atypicality. Moreover, these findings suggest that ASD intonation might not be as monotonous as described in other studies since a higher variation influences the perceived expressivity in the intensity contours. Thus, loudness could measure the dynamic intonation of autistic speech production (Bone et al., 2015), especially in tasks where affective prosody is investigated (Hubbard et al., 2017).

However, many researchers report the problem of having opposite results on intensity in the literature of reference.

For instance, in Mohanta et al. (2020; Mohanta & Mittal, 2022), the authors reported higher values in ASD (Quigley et al., 2016; Filipe et al., 2014) but also lower (Scharfstein et al., 2011). Furthermore, in addition to a trend of papers that present lower intensity in autistic speech (Chevallier et al., 2011; Ochi et al., 2019), there is also a consistent number of papers that did not find statistically significant results at all (Diehl & Paul, 2012; 2013; Filipe et al., 2014; Grossman et al., 2010). Moreover, in many studies, the authors decided not to study the intensity levels to avoid the risk of obtaining unclear results (Bisson et al., 2014; Dahlgren et al., 2018).

This difference in the results could be caused by the most reported problems: the recording environment and the microphone's position. The first cause reflects a common problem while doing a clinical speech collection and analysis since it is crucial to do so in a comfortable space for the patient. For the microphone, it would be necessary that all the participants wear it simultaneously from their mouths to ensure that all the variations are due to the actual speech production.

For these reasons, we decided not to consider the results obtained as relevant for the present. However, further studies could solve these impediments by rethinking the data collection process based on these observations.

Spectral-related parameters. In the present study, we found two spectral-related parameters with statistically significant results: the means of the spectral flux and the one from the slope of all the segments (voiced and unvoiced). Regarding the spectral flux, we found a statistically significant difference between both groups under all features extracted: voiced and unvoiced segments and the general mean that depends on them. Unfortunately, we did not find many literature studies for these features in the same context. Therefore, we hypothesised that these results show a trend typical of autistic speech. For example, Haider et al. (2019) report that jitter, shimmer, and spectral flux are valuable features to measure speech instability.

Furthermore, using a speech sample from patients with dementia, the authors demonstrated that these features make the difference in higher accuracy levels between different classifiers. The same observation was done by Bonnet et al. (2011) regarding the spectral characteristics and the pitch values. However, we did not find any other relevant studies to confirm the importance of spectral features to detect ASD, but we found in Pokorny et al. (2017) the same trends as the present study. They also used eGeMAPS to extract the acoustic features and found results shared with ours. Indeed, the ten most significant features between ASD and TD groups are slope in the 0-500 Hz range of unvoiced segments and mean of the length of these and voiced segments.

Furthermore, Volkmar (2017) posits the difficulty in registering the voice volumes visible on the spectrum because of the trend in autistic speech of having a small number of voice volumes that are usually louder than the typical speech. This trend may reflect the impairments in indicating areas of emphasis and higher values in some parameters, such as the spectral slope and flux.

Further studies can clarify these results with more focused research on the spectral-related parameters and show whether the differences between the unvoiced and voiced segments are significant in the early detection of the disorder.

4.3.2 Classification of Speech Samples through Machine Learning Algorithms

The previous sections explained the methods and the results obtained by applying supervised machine learning methods to the feature selection applied to the dataset. The final aim of these implementations is to test whether these acoustic features are typical in the speech production of Italian-speaking children with ASD. Good results in the performances of the classifiers would confirm this assumption. Moreover, testing the effectiveness of the extraction represents a general evaluation of the feature set used (eGeMAPS) since it was proposed as a standard for clinical purposes in acoustic analysis (Eyben et al., 2015).

The best DT, RF, and SVM models reach a high accuracy value (83%), meaning that the feature selection implemented obtained good results on the classifiers. Moreover, the DT, KNN, and SVM reach an optimal value of recall (100%) that indicates the recognition as true of all the acoustic features in the ASD speech.

However, by comparing the metrics obtained by the four best models implemented, we can exclude KNN because it has a poor performance overall, presenting overfitting between the train and the test sets.

Concerning RF, it had a decent performance without overfitting, and it is the model that reaches the highest level of precision (100%). Furthermore, it has the same values on accuracy as DT and SVM. However, these consistently outclassed RF for the other metrics, especially for the recall (50%).

Between the DTs and the SVMs models, if we only look at Table 3, the first can be selected as the best classifier on this dataset. Moreover, it is the only one reaching a recall of 100% on the test set of all the models trained (Table 4). However, if we compare the results of all the models obtained by the k-folds average on the test set, it likely overfits more than RF and SVM. Half of the ten best models of all the ones trained for DT are good in the classification task, but the others tend to decrease their performances drastically from the train to the test sets.

On the other hand, SVM reached high performances on almost all the evaluation metrics of the models trained. (Accuracy = 83% in eight models, Recall = 100%, Precision = 67% in nine models, F1-score = 80% in eight models, AUC = 88% in seven models out of ten).

In the literature, we found the same trend in the implementation of supervised classifiers on the features extracted by eGeMAPS (Asgari & Shafran, 2018; Lee et al., 2020; Li et al., 2019; Pokorny et al., 2017; Rybner et al., 2022; Schmitt et al., 2016). Shahin et al. (2019) used the same SVM model and described it as the most performant compared to GeMAPS, the previous features set. The linear kernel on SVM was also used in Li et al. (2019), and the authors aimed to reach high performances since this supervised method is the best to use when dealing with small datasets.

5. Conclusion

The present work analysed the speech production of Italian speaking children with Autism Spectrum Disorders (ASD) compared with their peers with typical neurodevelopment (TD). Unfortunately, there are no other similar studies on Italian compared with other languages to the best of our knowledge.

The main aim of this study was to determine whether the features reflected in both the qualitative and quantitative literature for English and other languages are also relevant for autistic production in Italian. Therefore, we performed an acoustic analysis on a specific dataset and implemented different types of supervised machine learning methods.

Our findings show that in the speech of Italian children with ASD, some typical acoustic features can be extracted and analysed as previously done in other languages. Furthermore, this task can be done using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) by considering the small sizes of the dataset used. However, further studies need to consider a larger collection of data and compare the performance of different feature subsets. In this way, it would be possible to create a standard set of typical acoustic features for children with ASD. The next step will be the ideation of a tool from a classifier able to distinguish the typical productions of the disorder from the not typical ones.

Further studies can analyse pitch and intensity features by paying attention to the recording process to satisfy all the requirements. However, due to the necessity of maintaining some environmental comforts for the patients in a clinical condition, we assume it is important to rethink the recording process to satisfy these requirements and collect audio data. Moreover, as pointed out by De La Fuente et al. (2020), the studies should use the same feature set to conduct the feature extraction to have the possibility to better compare the results between different languages.

To conclude, most of the problems found in this work concern the quality of recordings and the dataset size. However, the results obtained on the features extraction and classification are promising for developing a tool that can help the clinician diagnose the disorders at a young age.

6. Acknowledgements

The authors are deeply grateful to Sara Biancalani who collected the clinical data. The precious help of Annalisa Raffone and Ricardo Muñoz Sánchez is also acknowledged.

7. Bibliographical References

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, fifth ed.* (DSM-5). Washington (D.C.) / London: American Psychiatric Publishing.

Anolli, L. (2002). *Le emozioni*. Milano: Unicopoli.

Asgari, M., & Shafran, I. (2018). Improvements to harmonic model for extracting better speech features in clinical applications. *Computer Speech & Language*, 47, 298-313.

Biancalani, S. (2019). *Aspetti soprasegmentali e non verbali nel Disturbo dello Spettro Autistico: uno studio pilota*. Thesis dissertation, University of Florence.

Bisson, J. I., Cosgrove, S., Lewis, C., & Roberts, N. P. (2015). Post-traumatic stress disorder. *Bmj*, 351, h6161.

Boddaert, N., Chabane, N., Belin, P., Bourgeois, M., Royer, V., Barthelemy, C., Mouren-Simeoni, M. C., Philippe, A., Bunelle, F., Samson, Y., & Zilbovicius, M. (2004). Perception of complex sounds in autism: abnormal auditory cortical processing in children. *American Journal of Psychiatry*, 161(11), 2117-2120.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9-10), 341-345.

Bone, D., Black, M. P., Lee, C. C., Williams, M. E., Levitt, P., Lee, S., & Narayanan, S. (2012). Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist. *Proceedings of Interspeech 2012*, 1043-1046.

Bone, D. Lee, C. C., Black, M. P., Williams, M. E., Lee, S., Levitt, P., & Narayanan, S. (2014). The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *Journal of Speech, Language, and Hearing Research*, 57(4), 162-1177.

Bone, D., Black, M. P., Ramakrishna, A., Grossman, R. B., & Narayanan, S. S. (2015). Acoustic-prosodic correlates of 'awkward' prosody in story retellings from adolescents with autism. *Interspeech*, 616-1620.

Bonneh, Y. S., Levanon, Y., Dean-Pardo, O., Lossos, L., & Adini, Y. (2011). Abnormal speech spectrum and increased pitch variability in young autistic children. *Frontiers in human neuroscience*, 4(237), 1-7.

Chevallier, C., Noveck, I., Happé, F., & Wilson, D. (2011). What's in a voice? Prosody as a test case for the Theory of Mind account of autism. *Neuropsychologia*, 49(3), 507-517.

Chiew, J., Kjelgaard, M., Chiew, J., & Kjelgaard, M. (2017). The perception of affective prosody in children with autism spectrum disorders and typical peers. *Clinical Archives of Communication Disorders*, 2(2), 128-141.

Cho, S., Liberman, M., Ryant, N., Cola, M., Schultz, R. T., & Parish-Morris, J. (2019). Automatic Detection of Autism Spectrum Disorder in Children Using Acoustic and Text Features from Brief Natural Conversations. *Proceedings of Interspeech 2019,* 2513-2517.

Chollet, F. (2021). *Deep learning with Python*. Shelter Island (NY): Manning Publications Co.

Crystal, D. (2009). Persevering with prosody. *International Journal of Speech-Language Pathology,* 11(4), 257.

Dahlgren, S., Sandberg, A., D., Strömbergsson, S. Wenhov, L. Råstam, M., & Nettelbladt, U. (2018). Prosodic traits in speech produced by children with autism spectrum disorders–Perceptual and acoustic measurements. *Autism & Developmental Language Impairments*, 3, 1-10.

De La Fuente Garcia, S., Ritchie, C. W., & Luz, S. (2020). Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. *Journal of Alzheimer's disease. JAD*, 78(4), 1547-1574.

Diehl, J. J., & Paul, R. (2012). Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders. *Research on Autism Spectrum Disorder*, 6(1), 123-134.

Diehl, J. J., & Paul, R. (2013). Acoustic and perceptual measurements of prosody production on the profiling elements of prosodic systems in children by children with autism spectrum disorders. *Applied Psycholinguistics*, 34(1), 135-161.

Downey, A. (2012). *Think python. 2.0*. Needham (MA): Green Tea Press.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2), 190-202.

Filipe, M. G., Frota, S. Castro, S. L., & Vicente, S. G. (2014). Atypical Prosody in Asperger Syndrome: Perceptual and Acoustic Measurements. *Journal of Autism and Developmental Disorders*, 44, 1972-1981.

Grossman, R. B., Bemis, R. H., Skwerer, D. P., & Tager-Flusberg, H. (2010). Lexical and affective prosody in children with high-functioning autism. *Journal of Speech, Language, and Hearing Research*, 53(3), 778-793.

Haider, F., De La Fuente, S., & Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *Journal of Selected Topics in Signal Processing*, 14(2), 272-281.

Halberstam, B. (2004). Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *ORL*; *Journal for oto-rhino-laryngology and its related specialties*, 66(2), 70-73.

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4), 769-778.

Hubbard, D. J., Faso, D. J., Assmann, P. F., & Sasson, N. J. (2017). Production and perception of emotional prosody by adults with autism spectrum disorder. *Autism Research*, 10(12), 1991-2001.

Julião, M., Abad, A., & Moniz, H. (2020). Comparison of Heterogeneous Feature Sets for Intonation Verification. Proceedings of *International Conference on Computational Processing of the Portuguese Language*, 13-22.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child,* 2(3), 217-250.

Kim, J. C., Azzi, P., Jeon, M., Howard, A. M., & Park, C. H. (2017). Audio-based emotion estimation for interactive robotic therapy for children with autism spectrum disorder. *The 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, 39-44.

Kiss, G., Santen, J. P. V., Prud'Hommeaux, E., & Black, L. M. (2012). Quantitative analysis of pitch in speech of children with neurodevelopmental disorders. *13th Annual Conference of the International Speech Communication Association*, vol.2, 1342-1345.

Kissine, M., & Geelhand, P. (2019). Brief report: Acoustic evidence for increased articulatory stability in the speech of adults with autism spectrum disorder. *Journal of autism & developmental disorders*, 49(6), 2572-2580.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J.B., Grout, J., Corlay, S., & Ivanov, P. (2016). Jupyter Notebooks - A publishing format for reproducible computational workflows. In Loizides, F., Schmidt, B. (Eds.), *Positioning and Power in Academic Publishing: Players, Agents, and Agendas. Proceedings of the 20th International Conference on Electronic Publishing*, Amsterdam: IOS, 87-90.

Lee, H. Y., Hu, T. Y., Jing, H., Chang, Y. F., Tsao, Y., Kao, Y. C., & Pao, T. L. (2013). Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition. *Proceedings of Interspeech 2013*, 215-219.

Lee, J. H., Lee, G. W., Bong, G., Yoo, H. J., & Kim, H. K. (2020). Deep-learning-based detection of infants with autism spectrum disorder using auto-encoder feature representation. *Sensors*, 20(23), 6762.

Li, M., Tang, D., Zeng, J., Zhou, T., Zhu, H., Chen, B., & Zou, X. (2019). An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. *Computer Speech & Language*, 56, 80-94.

Luz S., Haider F., de la Fuente S., Fromm D., & MacWhinney, B. (2020). Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. *Proceedings of Interspeech 2020*, 2172–2176.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of mathematical statistics*, 18(1), 50-60.

Marchi, E., Schuller, B., Baron-Cohen, S., Golan, O., Bölte, S., Arora, P., & Häb-Umbach, R. (2015). Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages. *Proceedings of Interspeech 2015*, 115-119.

Mcallister, A., Sundberg, J., & Hibi, S. R. (1998). Acoustic measurements and perceptual evaluation of hoarseness in children's voices. *Logopedics Phoniatrics Vocology*, 23(1), 27-38.

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

Memari, N., Abdollahi, S., Khodabakhsh, S., Rezaei, S., & Moghbel, M. (2020). Speech analysis with deep learning to determine speech therapy for learning difficulties. *International Conference on Intelligent and Fuzzy Systems*, Springer, 1164-1171.

Mohanta, A., Mukherjee, P., & Mirtal, V.K. (2020). Acoustic Features Characterization of Autism Speech for Automated Detection and Classification. *National Conference on Communications (NCC)*, 1-6.

Mohanta, A., & Mittal, V. K. (2022). Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features. *Computer Speech & Language*, 72, 101287.

Nayak, V., Deshmukh, R., & Waghmare, S. (2019). Pitch pattern analysis in speech of children with autism spectrum disorder. *Journal of Innovative Technology Exploring Engineering,* 9(1), 4209-4212.

Ochi, K., Ono, N., Owada, K., Kojima, M., Kuroda, M, Sagayama, S., & Yamasue, H. (2019). Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder. *PLOS ONE,* 14(12).

Olivati, A. G., Assumpção, F. B., & Misquiatti, A. R. N. (2017). Acoustic analysis of speech intonation pattern of individuals with Autism Spectrum Disorders. *CoDAS*, 29(2).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., & Duchesnay, M. P. E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Pokorny, F., Schuller, B., Marschik, P., Brueckner, R., Nyström, P., Cummins, N., Bölte, S., Einspieler, C., & Falck-Ytter, T. (2017). Earlier Identification of Children with Autism Spectrum Disorder: An Automatic Vocalisation-Based Approach. *Proceedings of Interspeech 2017*, 309-313.

Quigley, J., McNally, S., & Lawson, S. (2016). Prosodic patterns in interaction of low-risk and at-risk-of-autism spectrum disorders infants and their mothers at 12 and 18 months. *Language Learning and Development*, 12(3), 295-310.

Rybner, A., Jessen, E. T., Damsgraad Mortensen, M., Larsen, S. N., Grossman, R., Bilenberg, N., Cantio, C., Jepsen, J. R. M., Weed, E., Simonsen, A., & Fusaroli, R. (2021). Vocal markers of Autism Spectrum Disorder: Assessing the generalizability of machine learning models. *Autism Research,* 1-13.

Ringeval, F., Marchi, E., Grossard, C., Xavier, J., Chetouani, M., Cohen, D., & Schuller, B. (2016). Automatic Analysis of Typical and Atypical Encoding of Spontaneous Emotion. *Proceedings of Interspeech 2016,* 1210-1214.

Robin, J., Harrison, J. E., Kaufman, L. D., Rudzicz, F., Simpson, W., & Yancheva, M. (2020). Evaluation of speech-based digital biomarkers: review and recommendations. *Digital Biomarkers*, 4(3), 99-108.

Rosenhall, U., Nordin, V., Sandström, M., Ahlsén, G., & Gillberg, C. (1999). Autism and hearing loss. *Journal of autism and developmental disorders*, 29(5), 349-357.

Sharda, M., Subhadra, T. P., Sahay, S., Nagaraja, C., Singh, L., Mishra, R., Sen, A., Singhal, N., Erickson, D., & Singh, N. C. (2010). Sounds of melody. Pitch patterns of speech in autism. *Neuroscience letters*, 478(1), 42-45.

Searle, J. R. S., & Vanderveken, D. (1985). *Foundations of illocutionary logic*. Cambridge: Cambridge University Press.

Shahin, M., Ahmed, B., Smith, D. V., Duenser, A., & Epps, J. (2019). Automatic Screening of Children with Speech Sound Disorders Using Paralinguistic Features. *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1-5.

Scharfstein, L. A., Beidel, D. C., Sims, V. K., & Finnell, L. R. (2011). Social skills deficits and vocal characteristics of children with social phobia or Asperger's disorder: A comparative study. *Journal of abnormal child psychology*, 39(6), 865–875.

Schelinski, S., & von Kriegstein, K. (2019). The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development. *Journal of Autism and Developmental Disorders*, 49(1), 68-82.

Schmitt, M., Marchi, E., Ringeval, F., & Schuller, B. (2016). Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices. *Speech Communication. ITG Symposium*, 12, 1-5.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., & Kim, S. (2013). The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Proceedings of Interspeech 2013,* 148-152.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., & Evanini, K. (2016). The Interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. *Proceedings of, Interspeech 2016,* 2001-2005.

Sharda, M., Subhadra, T. P., Sahay, S., Nagaraja, C., Singh, L., Mishra, R., Amit, S., Singhal, N., Erickson, D., & Singh, N. C. (2010). Sounds of melody. Pitch patterns of speech in autism. *Neuroscience letters*, 478(1), 42-45.

Shriberg, L. D., Paul, R., McSweeny, J. L., Klin, A., Cohen, D. J., & Volkmar, F. R. (2001). Speech and Prosody Characteristics of Adolescents and Adults with High-Functioning Autism and Asperger Syndrome. *Journal of Speech, Language, and Hearing Research*, 44(5), 1097-1115.

Styler, W. (2021). *Using Praat* for *Linguistic Research.* Version*: 1.8.3.* Last Update*:* March 12*,* 2021.

Tanaka, H., Sakti, S., Neubig, G., Toda, T., & Nakamura, S. (2014). Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative. *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 88-96.

Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2011). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of the Royal Society Interface*, 8(59), 842-855.

Van Rossum, G., & Drake, F. L. (2011). *The Python Language Reference Manual*. United Kingdom: Network Theory Limited.

Van Santen, J. P., Prud'Hommeaux, E. T., Black, L. M., & Mitchell, M. (2010). Computational prosodic markers for autism. *Autism*, 14(3), 215-236.

Volkmar, F. R., & Wiesner, L. A. (2017). *Essential Clinical Guide to Understanding and Treating Autism*. United Kingdom: Wiley.

Wilcoxon, F. (1945). Some uses of statistics in plant pathology. *Biometrics Bulletin*, 1(4), 41-45.