Towards a cumulative science of vocal markers of autism: a cross-linguistic meta-analysis-based investigation of acoustic markers in American and Danish autistic children

Riccardo Fusaroli[1,2,3], Ruth Grossman[4], Niels Bilenberg[5], Cathriona Cantio[5,6], Jens Richardt Møllegaard Jepsen[7,8], Ethan Weed[2,9]

1) Cognitive Science, School of Communication and Culture, Aarhus University

2) Interacting Minds Center, School of Culture and Society, Aarhus University

3) Linguistic Data Consortium, University of Pennsylvania

4) Communication Sciences and Disorders, Emerson College

5) Child and Youth Psychiatry, University of Southern Denmark

6) Psychology, University of Southern Denmark

7) Child and Adolescent Mental Health Centre, Mental Health Services in the Capital Region of Denmark, Copenhagen, Denmark

8) Center for Neuropsychiatric Schizophrenia Research and Center for Clinical Intervention and Neuropsychiatric Schizophrenia Research, Mental Health Services in the Capital Region of Denmark, Copenhagen, Denmark.

9) Linguistics, School of Communication and Culture, Aarhus University

Acknowledgments

We are very grateful to the participants in our studies as well as to the researchers and clinical practitioners who in various guises supported the collection of the data we analyzed in this project. We wish to thank our funding sources: the Interacting Minds Centre seed fundings "Clinical Voices" and "Clinical voices in the wild"; and the Danish Independent Research Council collective project "The Puzzle of Danish". Riccardo Fusaroli has been a consultant for F. Hoffmann-La Roche on related but not overlapping topics. The other authors have no real or potential conflicts of interest that could have had influenced the research.

Abstract

Acoustic atypicalities in speech production are widely documented in Autism Spectrum Disorder (ASD) and argued to be both a potential factor in atypical social development and potential markers of clinical features. A recent meta-analysis highlighted shortcomings in the field, in particular small sample sizes and study heterogeneity (Fusaroli, Lambrechts, Bang, Bowler, & Gaigg, 2017). We showcase a cumulative yet self-correcting approach to prosody in ASD to overcome these issues.

We analyzed a cross-linguistic corpus of multiple speech productions in 77 autistic children and adolescents and 72 neurotypical ones (>1000 recordings in Danish and US English). We replicated findings of a minimal cross-linguistically reliable distinctive acoustic profile for ASD (higher pitch and longer pauses) with moderate effect sizes. We identified novel reliable differences between the two groups for normalized amplitude quotient, maxima dispersion quotient, and creakiness. However, all these differences were small, and there is likely no one acoustic profile characterizing all autistic individuals. We identified reliable relations of acoustic features with individual differences (age, gender), and clinical feature: speech rate and ADOS sub-scores (Communication, Social Interaction, and Restricted and Repetitive Behaviors).

Besides cumulatively building our understanding of acoustic atypicalities in ASD, the study concretely shows how to use systematic reviews and meta-analyses to guide follow-up studies, both in their design and their statistical inferences. We indicate future directions: larger and more diverse cross-linguistic datasets, taking heterogeneity seriously, use of previous findings as statistical priors, understanding of covariance between acoustic measures, reliance on machine learning procedures, and open science.

Lay Summary

Autistic individuals are reported to speak in distinctive ways. Distinctive vocal production can affect social interactions and social development and could represent a noninvasive way to support the assessment of ASD. We systematically checked whether acoustic atypicalities found in previous articles could be actually found across multiple recordings and two languages. We find a minimal acoustic profile of ASD: higher pitch, longer pauses, increased hoarseness and creakiness of the voice. However, there is much individual variability (by age, sex, language, and clinical characteristics). This suggests that the search for one common "autistic voice" might be naive and more fine-grained approaches are needed.

Keywords

Autism Spectrum Disorder, Voice, Speech, Cross-linguistic

1. Introduction

Atypical prosody and voice are commonly-reported aspects of the speech of people with autism, which has been characterized as flat, sing-songy, pedantic, hollow, inappropriate, hoarse or hyper-nasal (Asperger & Frith, 1991; Goldfarb, Braunstein, & Lorge, 1956; Kanner et al, 1943; Pronovost, Wakstein, & Wakstein, 1966; Simmons & Baltaxe, 1975). Indeed, distinctive prosody is part of the diagnostic criteria in the ICD-10 and in the ADOS-2 assessment for autism (Lord, Rutter, Dilavore, & Risi, 2008; WHO 1992) and is indicated as one of the earliest-appearing markers of a possible Autism Spectrum Disorder (ASD) diagnosis (Oller et al 2010). These vocal factors may play a role in the socio-communicative impairments associated with the disorder. In addition to potentially impeding effective communication of e.g., emotional content (Travis & Sigman, 1998), they also generate negative responses from neurotypical raters, even when hearing as little as 1 second of speech (Grossman 2015, Sasson et al., 2017). These negative first impressions may have long term effects, e.g., providing a less optimal scaffolding for socio-communicative development, or even increasing the risks of social withdrawal and anxiety (Fay & Schuler, 1980; Paul et al., 2005; Shriberg et al., 2001; Van Bourgondien & Woods, 1992; Warlaumont, Richards, Gilkerson, & Oller, 2014). Given their potential role in affecting social functioning and in assisting diagnostic and assessment processes, it is important to understand how these vocal atypicalities manifest themselves across autistic people and uncover their acoustic underpinnings. This is especially true if we want to assess whether and how assessment and intervention tools should be developed to target them.

It has been nearly 80 years since unusual prosody was first reported by Kanner in 1943, and there is a growing interest in finding markers of ASD and social functioning. Nevertheless, two reviews of the field show that we know remarkably little about the precise perceptual and acoustic properties differentiating the speech of autistic people from that of neurotypical peers. A review of the literature from 2003 concluded that "No study offers a large number of subjects, matched with neurotypical children or adults (controlled for linguistic and non-verbal abilities). If findings were consistent, small-scale studies would offer pointers, but as it is these do not inspire confidence" (McCann & Peppé, 2003:347). A more recent systematic review and meta-analysis (Fusaroli et al., 2017) concluded that some single acoustic features (pitch mean and variability) showed robust small to moderate differences between groups. However, the studies reviewed were noted to have small sample size, high heterogeneity in methods and features analyzed, voice quality features - highlighted as important by speech pathologists and speech processing research - had been largely neglected, and that there was a need for multivariate approaches to account for shared variance and interactions across features. In other words, there is a need for a more rigorously cumulative scientific approach to the understanding of vocal and prosodic atypicalities in ASD.

In this paper we develop such an approach. First, we rely on the most recent meta-analysis of the field to set up the analysis of two new data sets. We build on the recommendations there produced, and test the replicability of the meta-analytic results (Fusaroli et al 2018). Second, as a cumulative approach might critically increase the number of acoustic features to analyze we explore principled ways to reduce the feature-space to a meaningful and interpretable subset of features.

1.1. Towards a cumulative research approach

A very common approach to cumulative research is to perform systematic reviews to map the field, and meta-analyses of previous results to achieve a more robust estimate of the underlying phenomena, beyond the variability of single studies. As an example, of the 17 studies conducted between 2010 and 2016, 13 found that people with autism had a wider pitch range, while 4 studies found the opposite effect (Fusaroli et al., 2017). A meta-analysis can pool the data from the different studies and perform an overarching inference as to the underlying effect size, and even assess whether systematic variations in study design (e.g., monological vs. dialogic speech production) might explain the differences in effects between studies (Cumming, 2014; Parola et al., 2020; Weed & Fusaroli, 2020). A common critique of this approach is “garbage-in-garbage-out”: If the studies included are too diverse, biased, or methodologically problematic, the meta-analytic inference will also be unreliable, and potentially overestimate effect sizes (Lewis et al 2020; Open Science Collaboration, 2015). While a few different techniques have been developed to assess the heterogeneity between studies and potential publication biases (Dwan, Gamble, Williamson, & Kirkham, 2013), they are not a solution to the issue of more reliably estimating the true effect, and the critique remains valid. Systematic reviews and meta-analyses are invaluable to get a feel for the field and identify potential issues or directions for research, but they should always be taken with caution as the researchers have no control on the quality and biases of the studies reviewed. We therefore need to critically combine systematic assessments of the field with well-targeted replications and new studies.

1.1.1 Building on existing guidelines

Previous systematic reviews and meta-analyses can be used to identify current best practices, pitfalls and blindspots, and therefore develop *guidelines for new studies* (Gelman, Jakulin, Pittau, Su, & others, 2008; König & Schoot, 2018; Williams, Rast, & Bürkner, 2018). Indeed, Fusaroli et al. (2017) identified several key areas for improvement in investigating vocal atypicalities in ASD.

1.1.1.1 More attention to the heterogeneity of the disorder

Building on insights from Fusaroli et al (2017), we designed a new study based on two existing corpora of voice data, collected in the US and Denmark (Cantio et al, 2016; Grossman, Edelson, & Tager-Flusberg, 2013). The study involves a high degree of heterogeneity in its sample: two diverse languages (Danish and US English) and a larger than average sample: 77 autistic participants and 72 neurotypical (NT) participants, against a previous median sample size of 17. Further, the study involves repeated measures of voice (between 4 and 12 separate recordings per participant). For each participant we have demographic (age, biological sex, native language) and clinical features (ADOS total scores, as well as the following sub-scores: Communication, Social Interaction, and Restricted and Repetitive Behaviors).

1.1.1.2. More systematic use of acoustic features across studies

Second, Fusaroli et al (2017) noted that different studies measured different acoustic features with diverse methods, without any explicit concern about comparing across studies[1]. Within our sample we systematically extract the acoustic features identified in the recent meta-analysis by Fusaroli (2017). This includes measures of pitch (median and variability), and rhythm (speech rate, average syllable length, pause number per unit of time, and length)[2].

Further, clinicians variously describe autistic voices as hoarse, creaky, breathy, harsh or otherwise dysphonic (e.g., Baltaxe, 1981; Pronovost et al., 1966; Sheinkopf, Mundy, Oller, & Steffens, 2000). We therefore identified in the speech signal processing literature acoustic features thought to be related to these perceptual qualities, e.g., pertaining to the glottal or spectral domain, fully listed in the methods section, and in Table S1.

1.1.1.3. Interdependencies between features and feature selection

Third, expanding the acoustic features investigated will produce a non-trivial increase in the number of statistical analyses required, potentially inflating the risk of false positives. Further, acoustic features are likely to be related to each other, and therefore we should assess whether all the features investigated provide independent information, and whether it is really necessary to add more complex acoustic measures of voice quality to the more traditional prosodic measures. Broadly speaking, there are at least four main approaches to the problem of feature-space reduction: 1) theoretically-justified a priori decisions, 2) dimensionality reduction methods, 3) clustering techniques, and 4) outcome-based methods. Each of these is a potentially viable method for reducing the feature space, but each comes with trade-offs. We briefly discuss each of these in turn.

Theoretically-justified a priori feature selection is the simplest of these. Choosing features a priori has the advantage of being perhaps the most easily interpretable of all four

[1] Exploration is a necessary component of research, and one should not put standardization in front of it, to avoid getting stuck with suboptimal methods (e.g., Devezer et al 2019; Wurbel, 2000). However, it is just as important, especially when discussing markers of disorders, to assess whether the findings generalize to new samples and how different methods compare to each other (Rocca and Yarkoni, 2021).

[2] Note that we did not include intensity-based measures, because we deemed them unreliable, due to their strong dependence on distance from the microphone, movements, etc. (Barsties & De Bodt, 2015).

approaches: given that the features have been chosen on theoretically informed grounds, the framework for interpreting them is already there.

Dimensionality reduction methods, such as Principal Component Analysis (PCA) are a class of methods which involve the data-driven inference of latent variables underlying the actual features investigated. The goal is to identify a small number of variables which can account for the majority of the variance in the acoustic features (Pearson, 1901). Because PCA transforms the feature-space into a smaller number of inferred features and does not distinguish between shared and unique variance among the original features, the components identified may be difficult to interpret within a theoretically meaningful framework (Preacher & MacCallum, 2003).

Network modeling approaches conceive of features as nodes in a network, and represent the shared variance between them graphically as connections between the nodes. An advantage to network models is that they represent the relationships between the original variables graphically, making them easier to interpret. A variety of algorithms exist for identifying "communities" of related variables, thus facilitating dimensionality reduction.

A final approach is outcome-driven (or supervised) feature selection. This common machine learning approach aims at identifying the minimum set of features most effective in discriminating between groups (Huang, 2015; Smialowski et al., 2010) These algorithms evaluate features based on their correlations with previously labelled data (Sheikhpour et al., 2017). Because this approach selects features from the original data set, it can maintain a reasonable degree of interpretability, although these techniques can easily choose a combination of features that do not make obvious intuitive sense.

We wished to explore common principled means of reducing these often intercorrelated acoustic features to a smaller subset of features. Ideally, these should be features which are not only useful for modeling the voice in a predictive framework, but which are also easily generalizable and crucially clinically intuitive. We therefore chose to focus on dimensionality reduction and network analysis. We discarded a priori feature selection, although we hope that over time, cumulative and theory-driven studies will lead to an a priori set of features. We also set aside outcome-based methods as our primary interest here is in understanding the broader landscape of acoustic features associated with the speech of people with autism, and not optimizing for predictive power with our particular data set.

1.1.1.4. Open science practices

To further promote cumulative approaches, we also provide an open data set including demographic, clinical and acoustic features, and open scripts to reproduce our analysis on the current data set and replicate and extend our findings on future data sets (https://osf.io/gnhw4/?view_only=3e51ee6253d548eb836af23ed9d01d73m; see also Wilkinson, 2016 and Parish-Morris et al., 2016).

1.2. Hypotheses

Based on the systematic review and meta-analysis and on current meta-scientific knowledge on replicability of meta-analytic findings, we developed the following expectations.

1. We will replicate meta-analytic findings that autistic people have: higher pitch mean and variability; more frequent and longer pauses; no differences in speech rate and syllable length, compared to neurotypical participants.

    a. Effect sizes will be half to a third smaller than previous meta-analytic findings due to hard to correct publication bias issues (Kvarven et al., 2020, see Table 2 for effect sizes of meta-analytic findings).

2. At least some of the measures of voice quality will be different in autistic people compared to neurotypicals, with effect sizes comparable to prosodic measures.

3. We expect the acoustic profile of autistic voice to be affected by individual differences (vs. a unique profile of autistic voice). In particular, we expect effects to be different by gender, and age; and acoustic features to relate clinical features of ASD as measured by ADOS sub-scores. In particular, increased pitch mean and variability, and pause number will relate to increased sub-scores, plausibly Social and Communication.

In a more exploratory fashion, we investigate whether acoustic features share variance, thus suggesting a priori ways of reducing the number of features investigated.

2. Materials and Methods

2.1. Participants and recordings

We collected two Danish and US English data sets involving 77 autistic participants and 72 neurotypical (NT) participants, each recording several audios, for a total of 1074 unique recordings. The Danish dataset included 29 autistic participants and 38 NT participants, retelling stories (Memory for stories, Reynolds and Voress, 2007) and freely describing short videos (Abell et al, 2000). The dataset included 335 recordings of autistic individuals and 427 recordings from NT participants. The US English data set included 48 participants of ASD and 34 NT participants, retelling stories (Grossman et al 2013). The dataset included 178 recordings of autistic individuals and 134 recordings from NT participants. The recordings had been collected for other purposes and their content – but not acoustics – analyzed in published studies (Cantio et al 2016; Grossman et al 2013).

Table 1. Participant characteristics. Clinical symptoms severity was measures using the Autism Diagnostic Observation Schedule – Generic (ADOS, Lord et al 2000). Cognitive functions were measured using the WISC-III for the Danish data (Kaufman, 1994), and the Leiter-R (nonverbal IQ, Roid & Miller, 1997) and the Peabody Picture Vocabulary test (receptive vocabulary, Dunn & Dunn, 2007). Note that the age spans are not precisely overlapping in the two corpora, but this is not an issue for the following analyses, given the effects are tested separately in the two corpora.

Language	Group	Age (months)	Males/ Total N	ADOS – Mean (SD)	Cognitive function
US English	NT	160.24 (36.57)	27/38	NA	Verbal IQ 114.58 (16.91) Nonverbal IQ 113.84 (9.71)
US English	ASD	152.83 (36.46)	24/29	Total: 13.94 (5.80) Communication 3.42 (1.71) Social 8.71 (2.57) Repetitive 1.59 (2.03)	Verbal 107.55 (19.15) Nonverbal 104.64 (15.49)
Danish	NT	130.53 (15.79)	31/34	NA	Verbal 108.59 (18.22) Nonverbal 102.57 (16.30)
Danish	ASD	132.00 (17.13)	46/48	Total: 11.31 (3.03) Communication 2.85 (1.43)	Verbal 100.72 (19.02) Nonverbal 103.14 (18.62)

	Social 7.04 (1.84) Repetitive 0.15 (0.46)

All recordings were pre-processed to remove background noise and interviewer speech when present. 32 acoustic measures were extracted (see Table 2 and 3). A full description of the process and features is available in the Supplementary Materials – S1.

2.2. Statistical modeling

2.2.1. Differences by diagnostic group

To assess whether autistic participants differed from NT participants, we ran Bayesian multilevel Gaussian regression models with the standardized acoustic feature as outcome, group (ASD vs. NT) and language (Danish vs. US English) as predictors (separately assessing the effects within language), and varying effects by participant (separately by language and group). Further details on the implementation and on the priors used are presented in the Supplementary Materials – S2, S3 and S5. We reported the estimated difference by group in terms of mean difference separately by language (that is, by corpus), 95% Compatibility Intervals (CIs, indicating the probable range of difference, assuming the model is correct) and Evidence Ratio (ER, evidence in favor of the effect observed against alternative hypotheses). When ER was weak (below 3, that is, less than three times as much evidence for the effect as for alternative hypotheses), we also calculated the Evidence Ratio in favor of the null hypothesis. Note that given the standardization of the outcome variables, the effect size is equivalent to Cohen's d, that is, is expressed in units of standard deviations.

To evaluate the potential role of individual differences in biological sex (Male vs. Female) and age, we built additional models, one per each suggested moderator interacting with group separately in the two languages. Age was modeled in terms of years and scaled. We reported the model estimates for the interaction, including CIs and ERs.

Note that we report additional analyses in the appendix to assess the robustness of the findings: we repeat all analyses on audio segments of 6 seconds to control for recoding length, as well as use the meta-analytic findings as priors to compare the change in inference with the models using skeptical priors, see Tables S2, S4, S5, S6. The results generally support our main findings and we report in the manuscript only qualitative divergences.

2.2.2. Relations to clinical features

To analyze the relation of acoustic and clinical features (ADOS total, Communication, Social Interaction, Repetitive Behaviors scores) we built multilevel Bayesian linear regression models with the acoustic feature as outcome (rescaled on a 0-1 scale) and clinical features as ordinal predictors, on the ASD group only, separately by language and with varying effects by participant (separately by language). We selected only features that were highlighted by the meta-analysis, as associated with group differences (pitch median and variability, speech rate, pause number and length), or with clinical features (jitter, Harmonic to Noise Ratio).

We otherwise followed the procedure described in the previous paragraphs. Further details on the implementation and priors are available in the Supplementary Materials - S4 and S5. Note that given the rescaling of the outcome and predictor variables, the effect size is on the scale of Pearson's r.

2.3 Feature-space reduction

We used two methods to explore feature-space reduction: Principal Component Analysis (PCA), and a spin glass community detection algorithm on a network model. PCA's and network models were calculated separately data sets. See supplementary materials for details, in particular S8, as well as Figure S1 and S2.

The data analysis scripts are available in the article repository at Open Science Foundation (https://osf.io/gnhw4/?view_only=3e51ee6253d548eb836af23ed9d01d73), and further details on the software employed is available in the Supplementary Materials – S5.

3. Results

3.1. Analysis of group differences in acoustic features

3.1.1. Acoustic features with meta-analytic results

The detailed results and comparison to the meta-analysis are reported in Table 2, and Figure 1. The results generally supported our hypotheses. We mostly replicated meta-analytic findings across both data sets (H1). Autistic participants across languages tend to use higher pitch, as well as fewer and longer pauses, and showed no differences in syllable length. Perhaps unsurprisingly, the effect sizes in our data are often smaller than in the meta-analysis (H1a), except for length of pauses. We also observe evidence for the importance of individual and linguistic differences (H3). Only in US English did we see robust evidence of slower speech rate and only in Danish did we see increased pitch variability. Further, biological sex and age interact with the effects, albeit inconsistently so across languages.

The findings are maintained if using only 6 second clips of the audio recordings, with the exception of syllable length becoming credibly longer in ASD in both languages (see Supplementary Materials – S9, and Figure S3).

Table 2. Estimated standardized mean differences (ASD – NT) for the six acoustic measures present in the meta-analysis. The first column reports the main effect of the diagnostic group (across sex and age), respectively from the meta-analysis, for Danish and for US English. The second column indicates the interaction between the effect of diagnostic group and biological sex (Male – Female), that is, the difference in effect of group between the male and the female participants. The third column reports the interaction between the effect of diagnostic group and age, that is, the change in effect size as age increases of 1 standard deviation. ER indicates the evidence ratio for the difference, ER01 the evidence ratio for the no effects. See Table S2 for a comparative perspective on the findings using skeptical (as here) and informed priors.

	Group (ASD - NT)	Biological sex (M - F)	Age
	β (95% CIs)	*β (95% CIs)*	*β (95% CIs)*
Pitch Median			
MA	0.38 (0.16 0.59)	NA	NA
DK	**0.12 (-0.08 0.32)**	**0.32 (-0.02 0.64)**	-0.01 (-0.06 0.03)
	ER = 5.26	**ER = 14.62**	ER = 2.34 ER01 = 9.73
US	**0.36 (0.12 0.61)**	-0.07 (-0.46 0.33)	**-0.02 (-0.06 0.02)**
	ER = 221	ER = 1.56 ER01 = 1.81	**ER = 4.13**
Pitch Variability			
MA	0.48 (0.26 0.7)	NA	NA
DK	**0.31 (0.16 0.46)**	**0.28 (-0.06 0.61)**	**0.01 (-0.04 0.05)**
	ER > 1000	**ER = 11.05**	**ER = 1.45 ER01 = 10.29**
US	0.02 (-0.2 0.23)	**-0.23 (-0.66 0.2)**	**-0.02 (-0.06 0.01)**
	ER = 1.3 ER01 = 2.32	**ER = 4.41**	**ER = 6.02**

Speech Rate			
MA	0.02 (-0.27 0.31)	NA	NA
DK	0.03 (-0.14 0.2)	-0.02 (-0.35 0.29)	0 (-0.04 0.04)
	ER = 1.55 ER01 = 2.71	ER = 1.19 ER01 = 2.15	ER = 1.12 ER01 = 11.28
US	**-0.11 (-0.28 0.05)**	**0.24 (-0.21 0.69)**	**-0.02 (-0.06 0.01)**
	ER = 6.74	**ER = 4.12**	**ER = 8.71**
Syllable Length			
MA	0.06 (-0.63 0.76)	NA	NA
DK	-0.02 (-0.14 0.09)	-0.03 (-0.3 0.25)	0 (-0.04 0.04)
	ER = 1.64 ER01 = 4.19	ER = 1.37 ER01 = 2.38	ER = 1.14 **ER01 = 13.57**
US	0 (-0.21 0.22)	-0.04 (-0.52 0.45)	0 (-0.03 0.04)
	ER = 1 ER01 = 2.16	ER = 1.28 ER01 = 1.48	ER = 1.1 **ER01 = 13.81**
Pause Number			
MA	0.4 (0.01 0.78)	NA	NA
DK	**-0.11 (-0.24 0.01)**	-0.05 (-0.34 0.22)	-0.01 (-0.05 0.03)
	ER = 14.04	ER = 1.6 ER01 = 2.55	ER = 2.36 **ER01 = 11.55**
US	**-0.17 (-0.35 0)**	**0.39 (-0.05 0.83)**	**-0.02 (-0.05 0.01)**
	ER = 17.6	**ER = 12**	**ER = 4.66**
Pause Length			
MA	0.21 (-0.09 0.5)	NA	NA
DK	**0.21 (0.03 0.39)**	**0.15 (-0.19 0.47)**	1.01 (-0.04 0.05)
	ER = 34.09	**ER = 3.4**	1.02 ER = 2.05 **ER01 = 10.31**
US	**0.27 (0.11 0.43)**	-0.12 (-0.55 0.31) E	0 (-0.03 0.03)
	ER = 332	R = 1.96 ER01 = 1.4	ER = 1.17 **ER01 = 16.13**

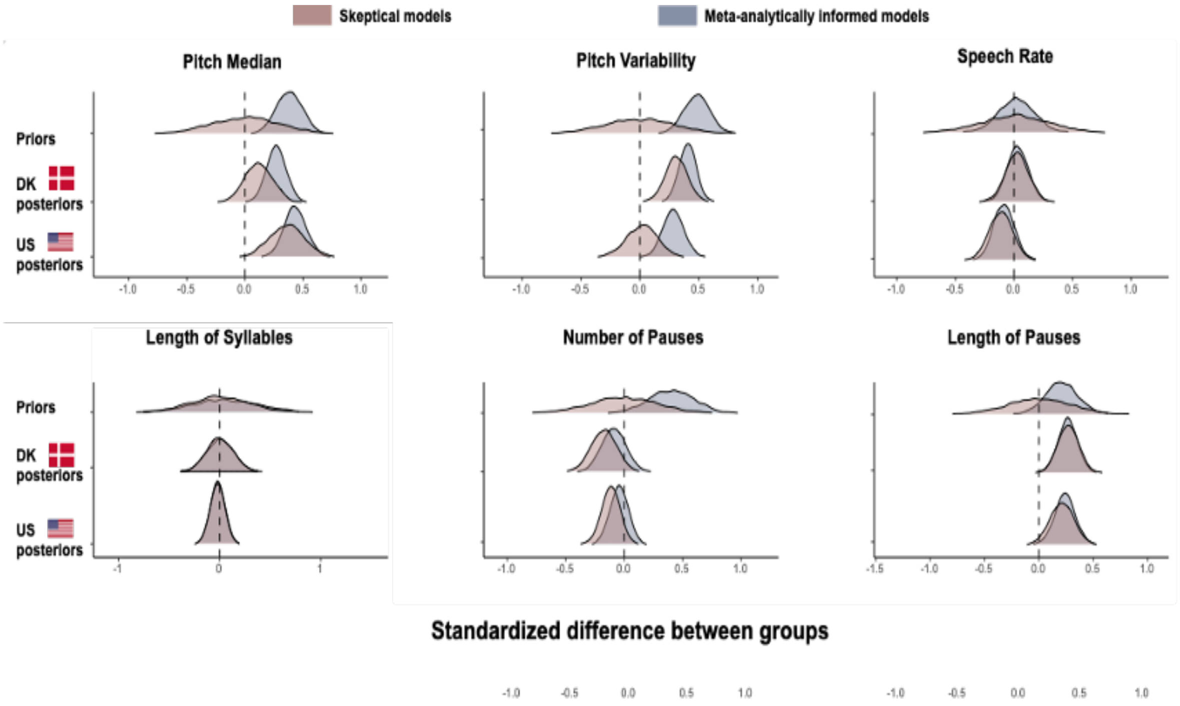

Figure 1: Comparing meta-analysis, skeptical expectations and results. Each panel presents a separate acoustic measure, with the x-axis corresponding to standardized mean differences (ASD - NT) equivalent to Hedges' g. The first row in each panel presents our prior expectation for effects: the skeptical expectations in red and the meta-analytic findings in blue. The second row represents the estimated difference (posterior) for Danish, and the third for US English. Estimated differences are in red for models using weakly skeptical priors (reported in the manuscript) and – for comparison - in blue for models using informed priors (reported in the appendix).

3.1.2. Novel acoustic features

The detailed results for each of the 26 features are reported in Table S3 in the appendix. We observe small to moderate (< 0.4) but reliable differences by group in the voice quality features within each data set, which are comparable to those in prosodic features (partially corroborating H2). As in more traditional acoustic features we see that including biological sex and age of the participants does in some cases affect the group differences (corroborating H3). However, strikingly, only three acoustic measures present the same small but reliable difference between the diagnostic group across the two languages (questioning the generalizability of H2). In particular, autistic participants have higher Normalized Amplitude Quotient (NAQ, effect sizes of 0.1 and 0.11), Maxima Dispersion Quotient (MDQ, effect sizes of 0.07 and 0.06) and creak (effect sizes of 0.13 and 0.15).

3.2. Relation with clinical features

Detailed results are presented in Table 3. While we can observe several reliable relations between acoustic and clinical features, the only consistent one across languages is speech rate (the slower the speech, the more severe the clinical feature), and to a lesser degree Harmonic to Noise Ratio (the lower, the more severe the clinical feature). Many correlations are small (< 0.2 or 4% of the variance), but some are moderate (between 0.4 and 0.54, that is, between 16% and 29% of the variance). The findings are analogous, albeit with smaller effect size in the 6 second audio recordings (see Supplementary Materials – S9).

Table 3. Estimated standardized relation between acoustic and clinical features. ER indicates the evidence ratio for the difference, ER01 the evidence ratio for the no effects.

	ADOS Total	*ADOS Communication*	*ADOS Social*	*ADOS Stereotyped*
	β (95% CIs)	*β (95% CIs)*	*β (95% CIs)*	*β (95% CIs)*
Pitch Median DK	0.09 (-0.39 0.62) ER = 1.51 **ER01 = 29.51**	0.09 (-0.2 0.4) ER = 2.4 **ER01 = 13.51**	0.06 (-0.37 0.52) ER = 1.41 **ER01 = 17.71**	-0.05 (-1.37 1.23) ER = 1.18 **ER01 = 3.4**
Pitch Median US	**0.17 (-0.17 0.52) ER = 3.68**	0.04 (-0.2 0.26) ER = 1.67 **ER01 = 17.41**	**-0.16 (-0.42 0.12) ER = 5.42**	**0.14 (-0.08 0.37) ER = 5.54**
Pitch IQR DK	0.02 (-0.3 0.32) ER = 1.26 ER01 = 48.74	**0.15 (-0.03 0.38) ER = 12.51**	-0.05 (-0.32 0.21) ER = 1.57 **ER01 = 28.93**	-0.1 (-1.21 0.88) ER = 1.41 **ER01 = 5.62**
Pitch IQR US	**0.17 (-0.09 0.45) ER = 5.83**	0 (-0.15 0.14) ER = 1.02 **ER01 = 28.49**	0.03 (-0.14 0.19) ER = 1.6 **ER01 = 37**	0.06 (-0.1 0.24) ER = 2.64 **ER01 = 16.66**
Speech Rate DK	**-0.18 (-0.42 0.01) ER = 14.87**	**-0.17 (-0.35 -0.04) ER = 87.89**	**-0.1 (-0.32 0.06) ER = 5.12**	**0.34 (-0.15 1.3) ER = 5.47**
Speech Rate US	**-0.07 (-0.2 0.05) ER = 4.6**	**-0.03 (-0.1 0.03) ER = 4.38**	**-0.05 (-0.12 0.02) ER = 8.66**	**-0.04 (-0.11 0.04) ER = 3.77**
Pause Number DK	**0.13 (-0.04 0.32) ER = 10.11**	**0.07 (-0.04 0.2) ER = 6.18**	**0.16 (0.02 0.37) ER = 33.19**	**-0.47 (-1.4 -0.03) ER = 30.5**
Pause Number US	-0.01 (-0.23 0.21) ER = 1.27 **ER01 = 70.95**	-0.03 (-0.14 0.08) ER = 1.97 **ER01 = 32.94**	**-0.06 (-0.2 0.07) ER = 3.51**	0 (-0.13 0.13) ER = 1.06 **ER01 = 24.89**
Pause Length DK	**-0.08 (-0.27 0.1) ER = 3.4**	**-0.07 (-0.21 0.05) ER = 5.32**	-0.03 (-0.19 0.13) ER = 1.68 **ER01 = 48.94**	0.18 (-0.53 1.16) ER = 2.27 **ER01 = 6.77**
Pause Length US	0.02 (-0.07 0.11) ER = 1.87 **ER01 = 148.46**	**0.03 (-0.02 0.08) ER = 5.16**	**-0.04 (-0.1 0.02) ER = 6.26**	0 (-0.05 0.06) ER = 1.07 **ER01 = 57.12**
Jitter DK	0.04 (-0.16 0.26) ER = 1.73 ER01 = 68.93	0.05 (-0.09 0.2) ER = 2.53 **ER01 = 28.63**	0.04 (-0.13 0.23) ER = 1.83 **ER01 = 43.24**	-0.2 (-1.08 0.43) ER = 2.51 **ER01 = 7.1**
Jitter US	**-0.06 (-0.15 0.04) ER = 4.97**	-0.02 (-0.08 0.04) ER = 2.67 **ER01 = 55.02**	0.02 (-0.05 0.09) ER = 2.07 **ER01 = 88.23**	-0.01 (-0.07 0.05) ER = 1.48 **ER01 = 50.57**

HNR Median DK	**-0.3 (-0.73 0.09) ER = 9.39**	**-0.24 (-0.54 0) ER = 18.32**	**-0.26 (-0.65 0.05) ER = 10.11**	**0.54 (-0.33 1.93) ER = 5.88**
HNR Median US	**-0.27 (-0.6 0.06) ER = 10.24**	-0.01 (-0.14 0.11) ER = 1.27 **ER01 = 34.04**	**-0.1 (-0.27 0.05) ER = 6.17**	**-0.21 (-0.53 0.06) ER = 6.84**
HNR IQR DK	**0.4 (-0.03 0.92) ER = 16.17**	**0.21 (-0.09 0.59) ER = 7.47**	**0.37 (0.03 0.85) ER = 27.78**	**-0.5 (-1.93 0.51) ER = 4.08**
HNR IQR US	-0.14 (-0.49 0.24) ER = 2.94 **ER01 = 30.82**	**-0.1 (-0.28 0.09) ER = 4.31**	**-0.27 (-0.47 -0.08) ER = 113.29**	**-0.15 (-0.35 0.06) ER = 7.55**

3.3. Feature-space reduction

Principal Component Analysis did not yield any insights into how the feature-space could be meaningfully reduced, see Figure 2. The first 10 principal components for each group cumulatively accounted for a substantial portion of the variance (DK NT = 87.8%, DK ASD = 87.7%, US NT = 93.5%, USA NT = 93.4%). However, the distribution of variance across the components was unequal between the two languages, with the first two components accounting for substantially more variance in the American English speakers than in the Danish speakers (see supplementary materials). Features from the three feature types were fairly evenly distributed along components 1 and 2, and no clear patterns were discernable. Both the Kaiser-Guttman rule and the scree plot method indicated that the feature space of the data could be reduced to approximately 7-9 components. Inspection of the relative contributions to the components did not suggest any clear pattern, although a full exploration of these data is beyond the scope of this paper. Scree plots and feature contributions can be found in supplementary materials.

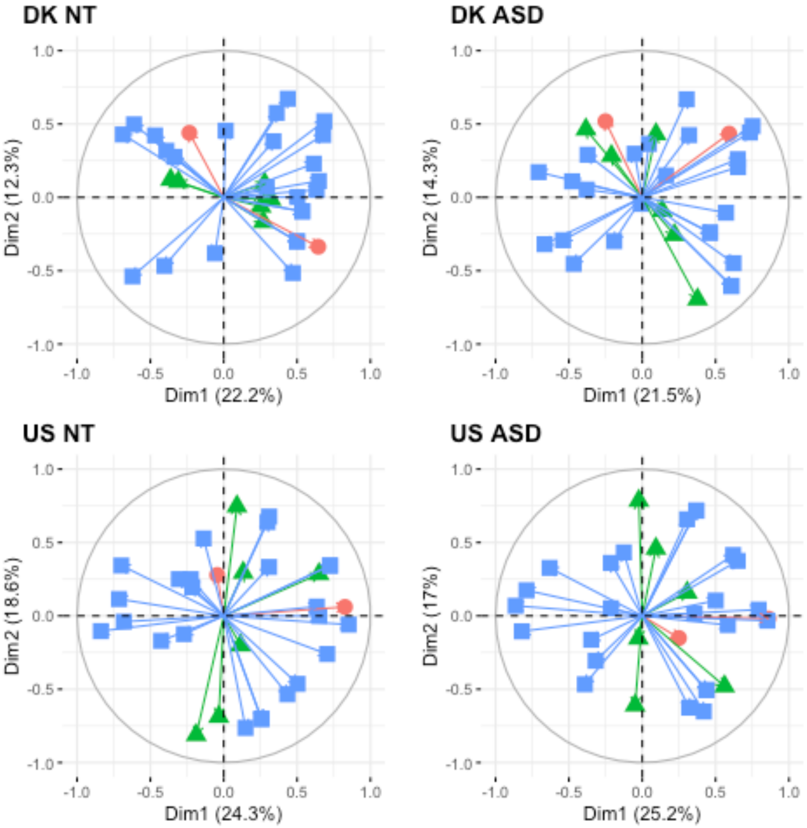

Figure 2: Features plotted against the first two principal components (Dimension 1 and Dimension 2). Green triangles represent rhythm features, red circles represent pitch features, and blue square represent voice quality features.

Spin-glass community detection did not yield any immediate insights into how the feature-space could be meaningfully reduced, either, see Figure 3. Three of the four groups settled on a three-community solution, while the fourth (DK NT) settled on a two-community solution. There was some indication that underlying patterns may exist, e.g., articulation rate, speech rate, and number of syllables were clustered in the same community in all four groups (see supplementary materials), however a full exploration of these data is beyond the scope of this paper.

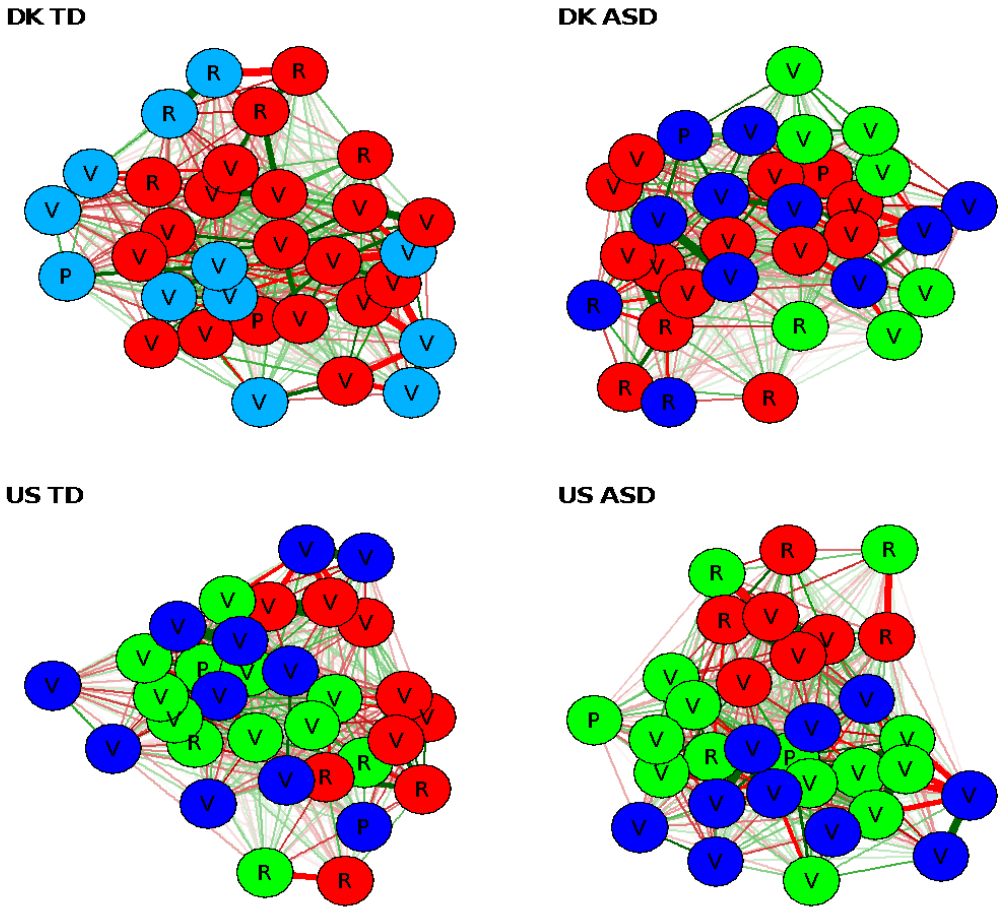

Figure 3: Partial correlation network graphs of acoustic features. Letters indicate the type of acoustic feature: "P" pertains to pitch; "R" to rhythm and duration; "V" to voice quality. Colors indicate feature communities identified with spinglass algorithm (Traag & Buggerman, 2009).

4. Discussion

In this work we aimed at building the foundations for a cumulative yet self-correcting approach to the study of prosody in ASD. Relying on a previous systematic review and meta-analysis of the field, we hypothesized that: H1) meta-analytic findings would replicate, potentially with smaller effect sizes; H2) voice quality measures would yield differences in the 2 groups analogous in size to those from prosodic measures; H3) individual demographic, clinical and linguistic differences would play an important role, defying the idea of a unique acoustic profile of ASD. We also set to explore whether the acoustic features showed obvious redundancies, allowing to reduce their numbers. In the following discussion we will consider how the findings bear on the hypotheses and explorations, highlight the limitations of the current study, and discuss further the cumulative yet self-correcting approach we propose.

We found a minimal (characterized by only few features) acoustic profile of ASD across Danish and US English: Autistic participants tended to use higher pitch, fewer but longer pauses, and increased NAQ, MDQ and creak compared to NT participants. Given the heterogeneity of previous studies and uncertainty about publication bias reported in the meta-analysis, these cross-linguistic replications (or lack thereof) are important. However, equally important is the focus that our findings put on linguistic, demographic, and clinical differences undermining the idea of a strong acoustic profile of ASD. There are many language-specific effects (e.g., speech rate being slower in ASD only for US English), and demographic differences (sex and age) affect even the cross-linguistically reliable acoustic features of ASD. From a clinical perspective, the moderate relations between acoustic measures and clinical symptoms are even more intriguing than diagnostic group differences. However, only speech rate and – to a lower extent - HNR show the same cross-linguistic relation: the slower the speech, and the lower the HNR, the more severe the clinical feature. Even more tellingly these features are not consistently different between diagnostic groups.

The findings thus suggest that there is no one general extensive acoustic profile of ASD. Systematic individual variations (sex, age, language, clinical features) should be always taken into account and we suspect that multiple clusters of acoustic profiles in ASD could be identified, all leading to the more general clinical descriptions of vocal atypicalities in ASD. However, to explore this idea and its potential clinical applications, we need to construct even larger cross-linguistic datasets systematically covering the heterogeneity in clinical features - and beyond - of autistic people, and more explicit normative modeling of individual variability (Marquand et al 2017).

Our exploration of feature reduction methods did not yield any clear finding. Future directions should explicitly include machine learning techniques targeting diagnostic group differences and relevant clinical features.

5. Conclusions

We set out to more cumulatively advance the study of acoustic markers in ASD, applying and assessing the recommendations and findings in a recent systematic review and meta-analysis. Across a relatively large cross-linguistic corpus, we identified a minimal acoustic profile of ASD (higher pitch, fewer and longer pauses, higher NAQ, MDQ and creak). However, we also highlight that individual differences in language, sex, age and clinical features relate to systematic variations in the acoustic properties of speech. This suggests that the search for a population-level marker might be naive and more fine-grained approaches are needed. We released the data and scripts used in the article to facilitate such future cumulative advances. The current study critically showcases a cumulative yet self-correcting approach, which we advocate should be more commonly used.

References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*, 1-16. doi: 10.1002/aur.174

Asperger, H., & Frith. (1991). *Translation and annotation of "autistic psychopathy" in childhood, by h. Asperger.* (U. Frith, Ed.). Cambridge, UK: Cambridge University Press.

Baltaxe, C. (1981). Acoustic characteristics of prosody in autism. *Frontier of Knowledge in Mental Retardation*, 223–233.

Barsties, B., & De Bodt, M. (2015). Assessment of voice quality: Current state-of-the-art. *Auris Nasus Larynx*, *42*(3), 183–188.

Cantio, C., Jepsen, J. R. M., Madsen, G. F., Bilenberg, N., & White, S. J. (2016). Exploring 'The autisms' at a cognitive level. *Autism Research*, *9*(12), 1328-1339.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49.

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. PloS one, 14(5), e0216125.

Dunn, L. M., & Dunn, D. M. (2007). Peabody picture vocabulary test–fourth edition (PPVT-4). *Circle Pines, MN: AGS*.

Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PloS One*, *8*(7), e66844.

Fay, W. H., & Schuler, A. L. (1980). *Emerging language in autistic children*. Hodder Education.

Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D., & Gaigg, S. (2017). Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis". *Autism Res*, *10*(3), 384–407.

Fusaroli, R., Weed, E., Lambrechts, A., Bowler, D., & Gaigg, S. (2018). Towards a cumulative science of prosody in asd. *INSAR 2018: Annual meeting*.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., & others. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Goldfarb, W., Braunstein, P., & Lorge, I. (1956). Childhood schizophrenia: Symposium, 1955: 5. A study of speech patterns in a group of schizophrenic children. *American Journal of Orthopsychiatry*, *26*(3), 544.

Grossman, RB (2015). Judgments of Social Awkwardness from Brief Exposure to Children with High-Functioning. *Autism*, 19(5):580-7.

Grossman, R. B., Edelson, L. R., & Tager-Flusberg, H. (2013). Emotional facial and vocal expressions during story retelling by children and adolescents with high-functioning autism. *Journal of Speech, Language, and Hearing Research*.

Huang, S. H. (2015). Supervised feature selection: A tutorial. *Artif. Intell. Research*, *4*(2), 22–37.

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

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. John Wiley & Sons.

König, C., & Schoot, R. van de. (2018). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review*, *70*(4), 486–509.
Lewis, M., Mathur, M., VanderWeele, T., & Frank, M. C. (2020). *The puzzling relationship between multi-lab replications and meta-analyses of the rest of the literature*.

Lord, C., Rutter, M., Dilavore, P. C., & Risi, S. (2008). *ADOS: Autism diagnostic observation schedule*. Hogrefe Boston, MA.

Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., & Beckmann, C. F. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular psychiatry*, *24*(10), 1415-1424.

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

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., ... & Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, *107*(30), 13354-13359.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Parish-Morris, J., Liberman, M., Ryant, N., Cieri, C., Bateman, L., Ferguson, E., & Schultz, R. T. (2016, June). Exploring autism spectrum disorders using HLT. In *Proceedings of the conference. Association for Computational Linguistics. Meeting* (Vol. 2016, p. 74). NIH Public Access.

Parola, A., Simonsen, A., Bliksted, V., & Fusaroli, R. (2020). Voice patterns in schizophrenia: A systematic review and bayesian meta-analysis. *Schizophrenia Research*, *216*, 24–40.

Paul, R., Shriberg, L. D., McSweeny, J., Cicchetti, D., Klin, A., & Volkmar, F. (2005). Brief report: Relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *35*(6), 861.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing tom swift's electric factor analysis machine. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *2*(1), 13–43.

Pronovost, W., Wakstein, M. P., & Wakstein, D. J. (1966). A longitudinal study of the speech behavior and language comprehension of fourteen children diagnosed atypical or autistic. *Exceptional Children*, *33*(1), 19–26.

Reynolds, C. R., & Voress, J. K. (2007). *Test of Memory and Learning (TOMAL 2)*. Austin, TX: Pro-Ed.

Rocca, R., & Yarkoni, T. (2020). Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction.

Roid, G. H., & Miller, L. J. (1997). Leiter international performance scale-revised (Leiter-R). *Wood Dale, IL: Stoelting*, *10*.

Sasson, N. J., Faso, D. J., Nugent, J., Lovell, S., Kennedy, D. P., & Grossman, R. B. (2017). Neurotypical peers are less willing to interact with those with autism based on thin slice judgments. *Scientific Reports*, *7*(1), 1–10.

Sheikhpour, R., Sarram, M. A., Gharaghani, S., & Chahooki, M. A. Z. (2017). A survey on semi-supervised feature selection methods. *Pattern Recognition*, *64*, 141–158.

Sheinkopf, S. J., Mundy, P., Oller, D. K., & Steffens, M. (2000). Vocal atypicalities of preverbal autistic children. *Journal of Autism and Developmental Disorders*, *30*(4), 345–354.

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*.

Simmons, J. Q., & Baltaxe, C. (1975). Language patterns of adolescent autistics. *Journal of Autism and Childhood Schizophrenia*, *5*(4), 333–351.

Smialowski, P., Frishman, D., & Kramer, S. (2010). Pitfalls of supervised feature selection. *Bioinformatics*, *26*(3), 440–443.

Traag, V. A., & Bruggeman, J. (2009). Community detection in networks with positive and negative links. *Physical Review E*, *80*(3), 036115.

Travis, L. L., & Sigman, M. (1998). Social deficits and interpersonal relationships in autism. *Mental Retardation and Developmental Disabilities Research Reviews*, *4*(2), 65–72.

Van Bourgondien, M. E., & Woods, A. V. (1992). Vocational possibilities for high-functioning adults with autism. In *High-functioning individuals with autism* (pp. 227–239). Springer.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological Science*, *25*(7), 1314–1324.

Weed, E., & Fusaroli, R. (2020). Acoustic measures of prosody in right-hemisphere damage: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, 1–14.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … others. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9.

Williams, D. R., Rast, P., & Bürkner, P.-C. (2018). *Bayesian meta-analysis with weakly informative prior distributions*.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Würbel, H. (2000). Behaviour and the standardization fallacy. Nature genetics, 26(3), 263-263.