Vocal markers of Autism Spectrum Disorder: assessing the generalizability of machine learning models

Astrid Rybner[1*], Emil Trenckner Jessen[1*], Marie Damsgaard Mortensen[1*], Stine Nyhus Larsen[1*], Ruth Grossman[2], Niels Bilenberg[3], Cathriona Cantio[3,4], Jens Richardt Møllegaard Jepsen[5,6], Ethan Weed[1,7], Arndis Simonsen[7,8,9], Riccardo Fusaroli[1,7,10]

N.B. * indicates shared first authorship.

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

2) Communication Sciences and Disorders, Emerson College

3) Child and Youth Psychiatry, University of Southern Denmark

4) Psychology, University of Southern Denmark

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

6) 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.

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

8) Psychosis Research Unit, Aarhus University Hospital

9) The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH.

10) Linguistic Data Consortium, University of Pennsylvania

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 funding "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 influenced the research.

Abstract

Background: Machine learning (ML) approaches show increasing promise to identify vocal markers of Autism Spectrum Disorder (ASD). Nonetheless, it is unclear to what extent such markers generalize to new speech samples collected in diverse settings such as using a different speech task or a different language.

Aim: In this paper, we systematically assess the generalizability of ML findings across a variety of contexts.

Methods: We re-train a promising published ML model of vocal markers of ASD on novel cross-linguistic datasets following a rigorous pipeline to minimize overfitting, including cross-validated training and ensemble models. We test the generalizability of the models by testing them on i) different participants from the same study, performing the same task; ii) the same participants, performing a different (but similar) task; iii) a different study with participants speaking a different language, performing the same type of task.

Results: While model performance is similar to previously published findings when trained and tested on data from the same study (out-of-sample performance), there is considerable variance between studies. Crucially, the models do not generalize well to new similar tasks and not at all to new languages. The ML pipeline is openly shared.

Conclusion: Generalizability of ML models of vocal markers - and more generally biobehavioral markers - of ASD is an issue. We outline three recommendations researchers could take in order to be more explicit about generalizability and improve it in future studies.

Keywords: *Autism Spectrum Disorder, voice, biobehavioral markers, machine learning, generalizability*

Introduction

Voice atypicalities are argued to be markers of Autism Spectrum Disorder (ASD) (McCann & Peppe, 2003; Fusaroli et al., 2017). Several machine learning (ML) studies have accordingly attempted to identify a clear acoustic profile of ASD, with promising results for supporting assessment processes (Hegde et al., 2019; Mohanta et al., 2020; Verde et al., 2018). However, any potential clinical application requires a careful assessment of how well the ML algorithms generalize not only to new participants from the same study sample, but also to new contexts, subpopulations and languages (Chekroud, 2018; Fusaroli et al., 2017; Low et al., 2020; Rocca & Yarkoni, 2020; Vandenbroucke et al., 2007; Yarkoni, 2020). The assessment of generalizability of ML models of vocal markers of ASD is still largely missing, and it is the focus of this study.

Currently, clinical features of ASD are predominantly assessed using the standardized protocols in Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule-Generic (ADOS-G), see (Lord et al., 2008). The process is time-consuming and resource-intensive, as it relies heavily on high quality training to ensure the validity of the examiners' ratings (Drimalla et al., 2020). Finding more automatically measurable markers could support, simplify and increase the reliability of the assessment process.

Vocalization patterns are associated with diverse cognitive, and motor abilities, as well as emotional states and levels of stress (Talkar et al., 2020; Williamson et al., 2017). Clinical features of ASD include repetitive behaviors, socialization and communication atypicalities, cognitive deficits, anxiety and sensory overload (Benson & Fletcher-Watson, 2011; Scheerer et al., 2020; Vargason et al., 2020), which are consequently likely to be reflected in autistic voice patterns (e.g., cognitive deficits resulting in increased pause behaviors; or anxiety resulting in increased jitter). Therefore, atypical vocalization patterns could serve as markers for these characteristics (Vargason et al., 2020; Yankowitz et al., 2019). While it is well accepted that autistic individuals often have atypical voices - e.g., sing-songy or monotone intonation and what has been referred to as inappropriate prosody (Baltaxe & Simmons, 1985; McCann & Peppé, 2003; Patel et al., 2019) -, acoustic investigations of the physical properties of voice underlying such perceptions often present weak or inconsistent findings. A recent meta-analysis identified increased and more variable fundamental frequency, as well as more and longer pauses as common characteristics (Fusaroli et al., 2017), these differences were, however, small and could only be partially replicated on new samples, across biological sexes, and languages (Fusaroli et al., 2018, 2021). Several ML studies have tried complementary approaches to these piecewise, highly top-down approaches. ML approaches can explore large numbers of features at once, and can therefore more fully rely on modern speech processing techniques and use low-level features that would be too numerous to deal with in traditional statistical frameworks. Further, more recent developments in neural networks can automatically construct the most relevant features to the task at hand from the raw speech signal (Badshah et al., 2017; Schneider et al., 2019; Sechidis et al., 2021). Capitalizing on this, ML models have been shown to be able to accurately differentiate previously unheard voice samples from new autistic and neurotypical (NT) individuals relying on vocal characteristics alone, with correct classifications ranging between 57 % and 96 % of the samples (Fusaroli et al., 2017).

However, ML algorithms are so apt at finding patterns that they often overfit to the data and learn distinctive patterns that are not present in datasets from other studies (Chekroud, 2018; Kuhn et al., 2013). For instance, one study (Bone et al., 2013) showed that algorithms with high performance in identifying autistic children in *The Interspeech 2013 Autism SubChallenge* dataset were relying on the different background noise and reverberation present in those specific recordings, more than on the actual voices. In other words, were the recording contexts to be switched between autistic and neurotypical children, or new recording contexts being used (e.g., the test being performed in a different school), the algorithms would likely perform at chance level. This example stresses the importance of ensuring the generalizability of machine learning models: in this context, the ability of the algorithms to perform accurately across different datasets. Generalizability has an impact on the inferences one can draw from a study: for instance, it can indicate whether the patterns found are related to an underlying biological atypicality in ASD (such as to motor control of the vocal cords) across all contexts and languages, or whether they are more specific to e.g., situations of high social pressure such as a conversation with a stranger. Generalizability has also an obvious impact on clinical applications, which require a clear delineation of how well algorithms might generalize to different recording contexts, subpopulations, and tasks before one can even consider real-world use. One should not only know whether an algorithm can be used as-is in a new context or with a new microphone, but also be aware of biases in assessments, e.g., of specific socio-demographic groups, which might have strong ethical and practical implications (Rocca & Yarkoni, 2021; Varoquaux & Cheplygina, 2021).

Little work has been done to examine the generalizability of the performance of ML models on vocal markers of neuropsychiatric conditions, including ASD (Fusaroli et al., 2017;

Low et al., 2020; Parola, Simonsen, Bliksted, Zhou, et al., 2020). Algorithms might provide a reliable performance (i.e., sensitivity and specificity in classifying data) only if the data keep certain properties constant: e.g., it is always recorded with a given type of microphones, or following specific procedures to minimize movement in the speaker, or using a certain task, or within a given language and no other. That is not in itself an issue, and can still lead to useful insights and applications. However, it is important to acknowledge such potential model limitations to avoid inappropriate interpretation of the results or overselling of practical applications.

This study specifically aimed to investigate the generalizability of ML models of vocal markers of ASD. We searched the literature to identify a promising ML model and replicated its training - within a highly conservative pipeline - on a 3-study cross-linguistic dataset. This allowed us to investigate the following three questions:

- Q1: How well do models generalize to different participants from the same study, performing the same task and speaking the same language?
- Q2: How well do models generalize to the same participants performing a different task (describing videos vs. repeating a story) in the same language?
- Q3: How well do models generalize to participants from a different study, performing the same kind of task (repeating a story) but speaking a different native language?

The scope of this paper is to test generalizability of ML models of vocal markers of ASD and hence, we identify three essential proposals for future research.

Methods

Model identification

To identify promising algorithms, we updated the literature search for ML studies of vocal markers of ASD reported in Fusaroli et al (2017), identifying 23 studies (for a more elaborate description of the procedure and a table summarizing the studies, see appendix A.1). We then selected one of the several algorithms that could be applied on relatively small samples with standard computational setups (e.g., trained Support Vector Machine (SVM) or random forest algorithms as opposed to deep neural networks), and transparently reported the methodological choices. We focused on relatively simple models as they tend to be less prone to overfitting, therefore if we were to find generalizability concerns (i.e., inability to accurately identify autistic participants in new samples), such concerns would be even more relevant for models like neural networks, more complex and more likely to overfit. We selected Shahin et al. (2019), a study based on SVM and rigorous cross-validation (CV), reaching high performance in predicting ASD from voice. It reported an accuracy of 0.88, that is, 88% of the samples were correctly classified; and a F1 score of 0.90. F1 scores are used to calculate performance when e.g., the groups to separate have a different number of cases, and is defined as the harmonic mean of precision (fraction of true positive examples among the examples that the model classified as positive) and recall (fraction of true positive examples among the examples that the model classified as positive). SVM is often used for binary classification problems, and it has been shown to perform better than e.g. neural networks and Naive Bayes when there is a limited amount of training data and a high number of features (Kirk, 2017), which is the case in the current study.

Pipeline

This study sets the methodological choices from Shahin et al. (2019) in a highly conservative and fully reported pipeline, which relies on cross-validated training procedures and held-out testing sets, that is, it ensures the model is trained (fitted) on a subset of the data (training dataset), but its performance is assessed on a subset of the data on which it has not been fitted (held-out dataset). Figure 1 provides an overview of the pipeline, which is discussed in detail below.

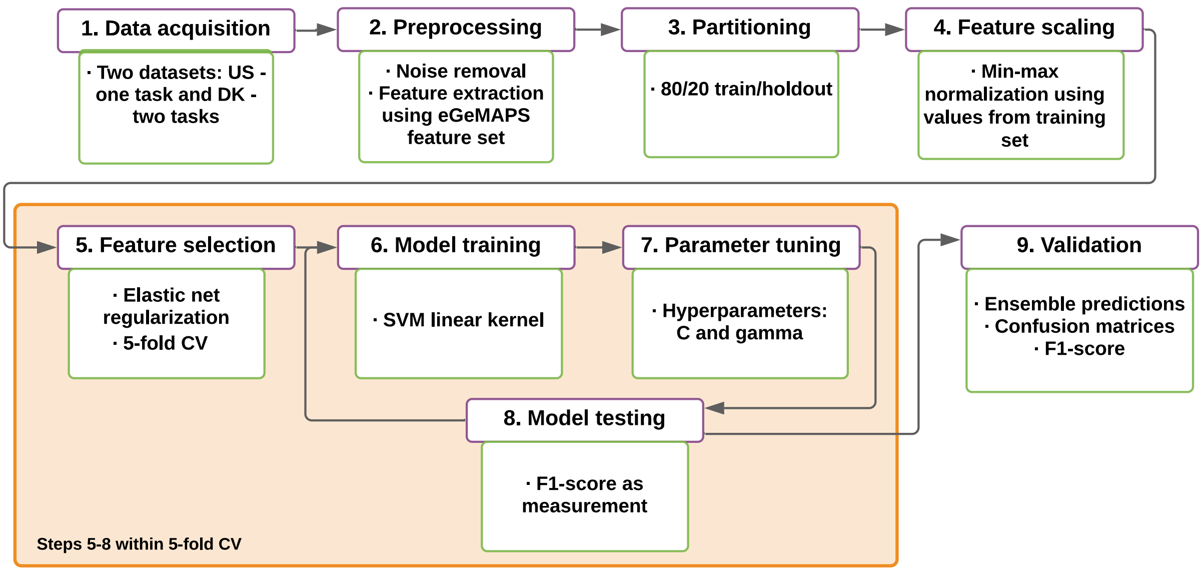

Figure 1. Overview of the machine learning pipeline. Purple refers to the general steps whereas green refers to specifics. "US" indicates the US English study, "DK" the two Danish studies. eGeMAPS indicates the Extended Geneva Minimalistic Acoustic Parameter Set (Eyben et al xxx). Train and holdout indicate respectively the portions of the dataset on which the model is fitted (trained) and tested. Elastic net indicates a regularization procedure used to reduce the number of features

included in the model. SVM RBF indicates the model fitted: Support Vector Machine with a Radial Basis Function. C and gamma parameter indicates respectively how much error is tolerated on the margin and how curved the margin is expected to be.

Data sources

The dataset used in this study consists of voice recordings collected in previous studies for other purposes and their content has been analyzed in other published research (Cantio et al., 2016; Fusaroli et al., 2021; Grossman et al., 2013). For demographic, clinical and cognitive information about the participants, see Table 1[1].

Table 1. Demographic and clinical features of the participants. DK stands for Danish samples, while US for US English samples. ADOS subscores are abbreviated: "Com." stands for Communication, "SI" for reciprocal Social Interaction, "Crea." for Creativity and "Stereo" for Stereotyped Repetitive Behaviors. IQ stands for intellectual quotient.

	Participants (recordings)	Age in months: mean (sd)	Mean (sd) ADOS scores				mean (sd) IQ scores	
			Com.	SI	Crea.	Stereo	Verbal	Nonverbal
DK	**51 (610)**	**133 (15.2)**	--	--	--	--	105 (19.4)	102 (15.9)
ASD	24 (288)	133 (16.5)	2.71 (1.52)	6.95 (1.80)	1.29 (0.56)	0.19 (0.51)	99.2 (19.2)	101 (18.1)
NT	27 (322)	133 (14.3)	--	--	--	--	110 (18.5)	102 (13.8)
US	**81 (309)**	**157 (36.8)**	--	--	--	--	109 (18.9)	109 (14.1)
ASD	50 (187)	153 (36.3)	3.46 (1.73)	8.9 (2.65)	1.06 (0.85)	1.73 (2.12)	105 (19)	106 (15.9)
NT	31 (122)	163 (37.4)	--	--	--	--	115 (17.8)	114 (9)

[1] Note that female participants' data in our original datasets were very sparse: e.g. the US dataset contains only 5 autistic girls (3.8%). Since exploratory analyses showed little ability of the models to generalize to female voices, but female participants were too few and unbalanced across datasets to be able to draw any conclusion, we excluded them from the current study, and therefore the table. For a more detailed analysis of differences in vocal markers of ASD depending on biological gender see Fusaroli et al. (2021).

We collected two Danish and US English datasets involving 74 autistic participants and 58 neurotypical (NT) participants, all with verbal and non-verbal cognitive function within a typical range. Each participant recorded several audios, for a total of 919 unique recordings. The Danish dataset included 24 autistic participants (288 recordings) and 27 NT participants (322 recordings), retelling stories (Memory for stories, Reynolds & Voress, 2007) and freely describing short videos (Abell et al., 2000). The US English dataset included 50 autistic (187 recordings) and 31 NT (122 recordings) participants, retelling stories (Grossman et al., 2013). The recordings had been collected for other purposes and analyzed with different purpose in previous studies (Cantio et al., 2016; Fusaroli et al., 2021; Grossman et al., 2013).

Note that the two samples are only roughly matched. While cognitive function and clinical features as measured by ADOS are largely overlapping, US participants are a bit older than Danish ones, and present a larger variability in age. Further, the language spoken, while always a Germanic one, is obviously different. In particular, Danish is often characterized as an atypical language with strong reduction in consonant pronunciation (Trecca et al., 2021), albeit no systematic comparison with US English has been performed to our knowledge. These differences are not a conceptual issue for the following analyses: it is crucial to assess whether so called vocal markers of ASD can generalize across corpora with different characteristics and explore how these differences might matter for the generalizability of the findings.

Preprocessing

The recording equipment and procedure were consistent within language, but not further specified in the original studies. Experimenter voice, background room noise, reverberation, and hum were removed from the audio recordings using iZotope RX 6 Elements (Izotope, 2017). Long-term average spectra for each recording were inspected for possible noise artefacts and further cleaned if any were found (Olsen, 2018). Following Shahin et al. (2019), we extracted the standardized voice feature set eGeMAPS with the open-source software OpenSmile (Eyben, 2015, p. 201; Eyben et al., 2010, 2016). The feature set contains 87 features, which are described in appendix A.2. All features in the training data were scaled using min-max normalization to achieve a dataset with a common scale without losing information or distorting differences in the range of values (Singh & Singh, 2020). The values used to normalize the training set were saved and used to normalize the test set, thus avoiding potential information leakage between the two sets.

Model training procedure

To investigate Q1, Q2 and Q3, three models were trained. The first model was trained on the Danish Animated Triangle data, the second on the Danish storytelling data, and the third on the US storytelling data. After training, the models were tested on different test sets. The combinations of training and test sets can be seen in figure 2.

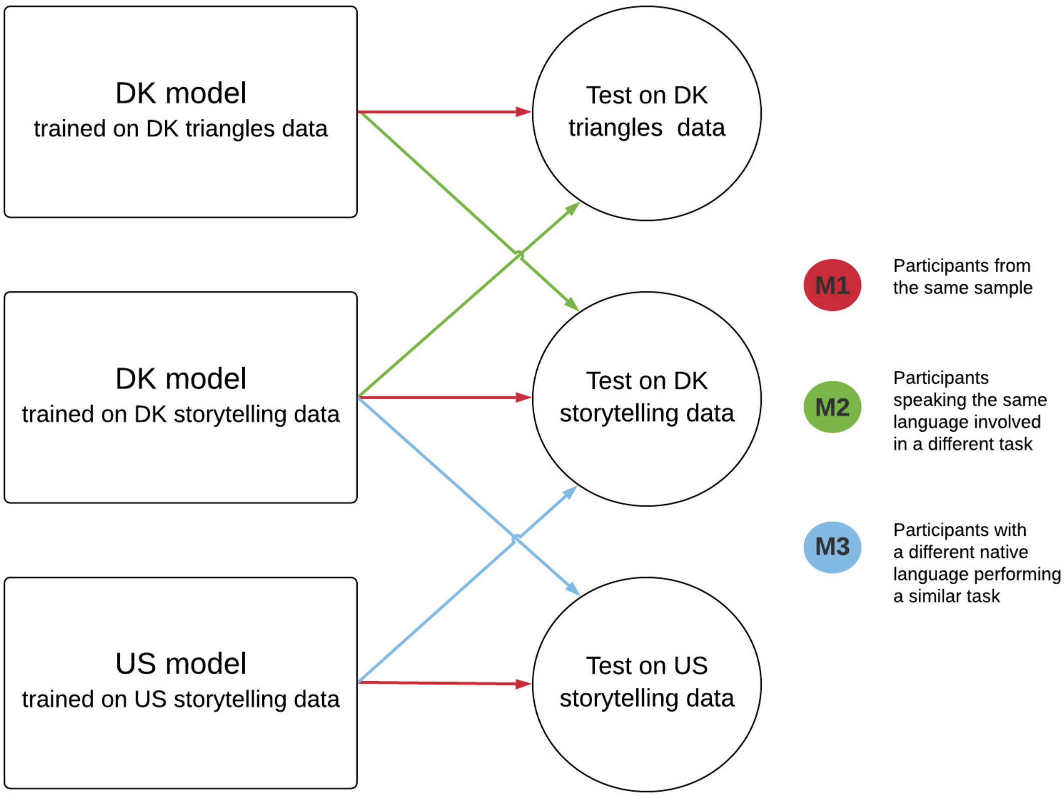

Figure 2. Overview of model training and testing. Each purple square indicates a 5-model voting ensemble trained on one dataset. Each beige circle indicates a held-out testing dataset. The arrows show which models are tested on which held-out dataset and their color-codes for the question that is being asked: Red for Q1 (new participants from the same sample), green for Q2 (same participants, different task), and blue for Q3 (same task, different language).

Each dataset was partitioned into a training set and a test set with an 80/20 split, stratified by participant ID; all data points for a given participant ended up in the same partition. We also ensured the test set included an equal number of participants from each diagnostic group (total number of voice recordings differed in some cases - see appendix A.5) to support more robust out-of-sample performance. This test set was used to answer questions about the generalizability to different participants from the same sample, performing the same task (Q1). When testing Q2 and Q3, the test dataset consists of either the same participants performing a different task (Q2) or participants with a different native language performing the same kind of task (Q3).

To prevent overfitting as much as possible, we relied on a strict five-fold cross-validation of the training process. In other words, the training dataset was split into five folds (roughly equally sized subsets) balanced by participant ID, so that recordings from the same participant only appeared in one fold and the testing fold only contained never seen before participants. Each of these five folds was then used as a validation set for a training set composed of the other four folds (see figure 1). Within each of these five CV training sets, we performed feature selection, and tune γ (gamma, or the curvature of the margin between classes) - and C (or the error tolerated at the margin) -hyperparameters using a grid search (see appendix A.3 for details), feature selection using elastic net and SVM classifiers with radial basis function (RBF) kernels using the Scikit-learn module in Python (Pedregosa et al., 2011; Van Rossum & Drake, 2009). Due to the cross-validation process, this yielded five trained SVM models for each of the three datasets used.

Majority voting

The 5 SVM models were used to assess the test sets, and their predictions were combined into a single voting ensemble (Brownlee, 2020; Hansen et al., 2021; Sechidis et al., 2021). Each model made a prediction for each voice recording in the test set, and the ensemble model gave a final predicted class based on the majority of these model votes. Note that other systems beyond majority rules have been developed, e.g., Mixture of Experts with weights based on similarity between test and training data (Hansen et al., 2021; Sechidis et al., 2021). Combining or utilizing multiple models within a single model – such as an ensemble model – benefits performance and generalizability, since no two models are likely to overfit in the same way and different models can compensate for each other's biases (Buracas & Albright, 1993; Hong & Page, 2004; Tang et al., 2005).

Software implementation and open science

All steps in the analysis – except for the de-noising – are implemented using open software. The feature extraction software toolkit openSMILE 3.0 was used to extract the feature sets GeMAPS, eGeMAPS and ComParE from the speech signals (Eyben et al., 2010). Data cleaning, including removal of outliers and normalization, and feature reduction using ElasticNet was implemented using R (RStudio Team, 2020). Model training for all kernels and SVMs as well as validation and testing was carried out using the Scikit-learn module in Python (Pedregosa et al., 2011; Van Rossum & Drake, 2009). The source code for the analysis can be found at: https://osf.io/9mtpk/

Results

This section presents the performance of the ML models when using various training and testing sets. A crude overview of the performance of the seven models (three testing on the same task in the same language, two testing on a different task in the same language, two testing on the same task in a different language) is given in Table 2. We include plots of precision (fraction of true positive examples among the examples that the model classified as positive) and recall (fraction of true positive examples among the examples that the model classified as positive) (Figure 3) on test sets to enable a more precise error analysis of the results. Model performance is evaluated using F1-score, the harmonic mean of precision and recall. Confusion matrices displaying the raw performance can be seen in Appendix A.4.

Tested on → Trained on ↓	DK triangles data	DK storytelling data	US storytelling data
DK triangles data	**M1a** - Q1 0.59	**M2a** - Q2 0.60	--
DK storytelling data	**M2b** - Q2 0.66	**M1b** - Q1 0.89	**M3a** - Q3 0.45

US storytelling data	--	**M3b** - Q3 0.40	**M1c** - Q1 0.64

Table 2 Overview of model performance on different test sets, F1-scores. The colors code the question tackled by the specific test set: M1 (red) answers question 1; M2 (green) answers question 2; M3 (blue) answers question 3.

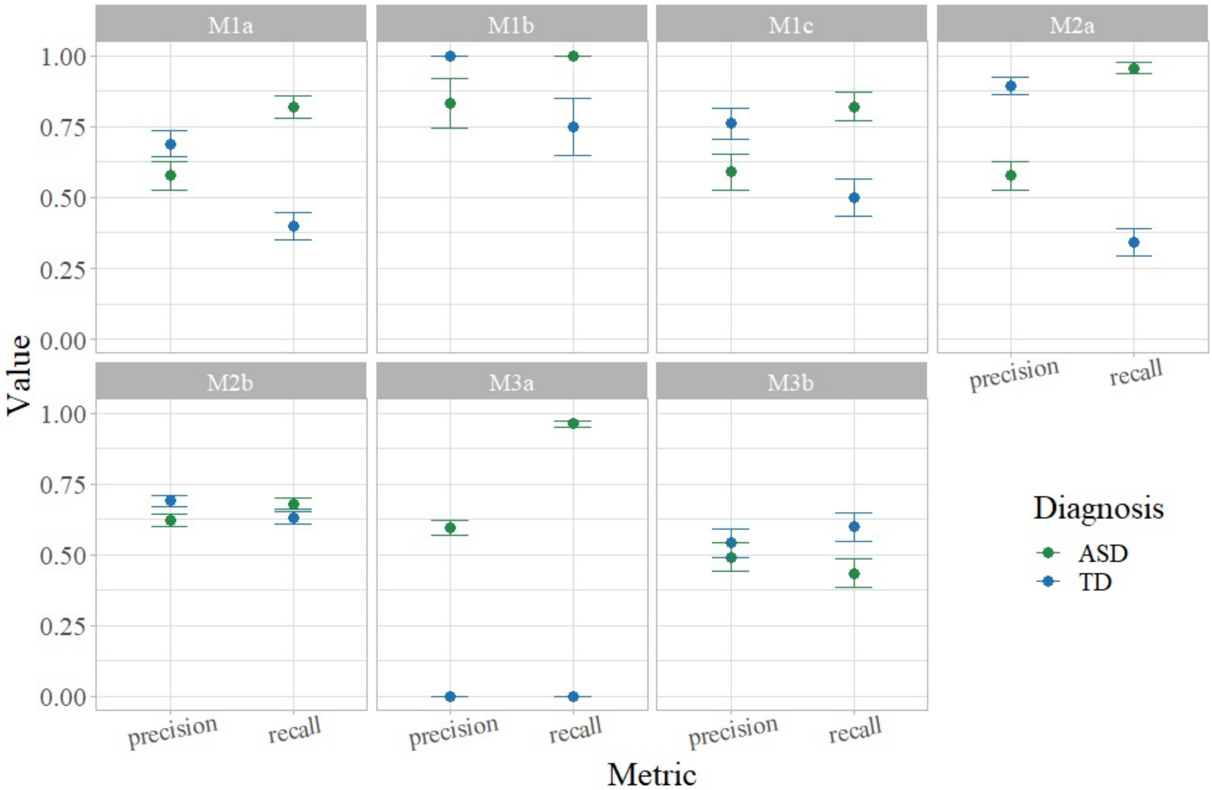

Figure 3: Precision (fraction of true positive examples among the examples that the model classified as positive) and recall (fraction of true positive examples among the examples that the model classified as positive) for ASD and NT, respectively, for all models. Compare with table 2 (overview of models)

Discussion

In this work we aimed at systematically assessing the generalizability of ML results on vocal markers of autism. We relied on cross-linguistic data to assess whether the results would generalize: (Q1) to different participants from the same study, performing the same task; (Q2) to a different task performed by the same participants; and (Q3) to participants from a different study, with a different native language performing the same type of task. We briefly discuss our specific findings in relation to these three questions. We then advance more general recommendations for how future ML studies could take generalizability concerns more seriously.

When assessing Q1, we found that all models, independently of the dataset they were trained on, were above chance at identifying autistic participants amongst new participants from the same study, with performance within the range of previously published articles: accuracies between 62% and 87% (Fusaroli et al., 2017). However, the very same training procedure yielded highly variable results on the three datasets (F1-scores between 0.59 and 0.89). Previous studies suggest that different demands of diverse tasks, used to collect speech recordings, do impact how clearly vocal atypicalities are expressed by participants (Fusaroli et al., 2017; Parola, Simonsen, Bliksted, & Fusaroli, 2020). In particular, more cognitively and socially demanding tasks tend to show bigger vocal differences between neurodiverse and neurotypical populations. However, we doubt that this is the explanation of our findings, since we do not see any reason to suggest that retelling stories in Danish is inherently more challenging than describing previously watched videos in Danish, or retelling stories in US English. Additionally, exploratory analyses on the very few female participants suggested little ability to generalize to new participants of another biological sex. These findings are worrying, since they indicate that even re-training ML models on one's own dataset is not guaranteed to yield the same performance as published findings.

When assessing Q2, we found again that the models were able to assess the same participants performing a new task with an accuracy above chance, but generally lower than on the original task (F1 scores of 0.60 and 0.66). The participants are the same, the tasks are relatively similar, and the conditions (and date) of recording are exactly the same, so we speculate that some overfitting to the specific task is at stake.

When assessing Q3, we found that the models were not able to assess other languages with the same type of task, with performance dropping below chance (F1 scores of 0.4 and 0.45). Relatedly, one previous study relied on a cross-linguistic corpus (English, Hebrew, Swedish and French, with no reported background noise removal procedure). They reported inconsistent results as to whether new participants speaking a different language could be identified above chance, and low performance even in that case (Schmitt et al., 2016). Transferring a ML model of autistic voice across languages is thus non-trivial, and we should not expect findings to generalize, without re-training of the model.

These findings raise strong concerns as to the generalizability of ML models of voice in autism. First, performance of the same model trained on relatively similar datasets changed quite drastically, indicating that evaluations of ML models on one dataset might not be representative of the results that could be achieved on a different dataset. Second, generalizability across relatively similar tasks (the description of a narrative video vs. the retelling of a story) is low, even when keeping the participants and native language constant. Third, even when testing phylogenetically close languages (US English and Danish being both Germanic languages), the acoustic patterns found by our models could not generalize. One could speculate that testing on a Romance language (e.g., Italian), or a non-Indo-European language (e.g., Mandarin Chinese) would compromise generalizability even further.

In summary, while models can be trained on novel datasets to reach roughly comparable performance to original reports, there is considerable variance, and these levels of performance do not generalize beyond the specific task and language. Since these issues emerge already with relatively simple SVM classifiers, with a limited amount of hyperparameters to set, we expect them to be more prominent in more complex model architectures such as convolutional neural networks or other deep learning methods, as these models are known to be more prone to overfitting (Kuhn & Johnson, 2019). Since the study of autistic voice is generally constrained by limited data availability (Fusaroli et al., 2017), the issue is even more prominent. We therefore argue that it is crucial to explicitly tackle the issue of generalizability in order to advance the field.

Moving beyond the specific case of our own datasets and ML procedures, we want to make a more general perspective for taking generalizability concerns more seriously in ML research, and provide the following three key guidelines that we elaborate on below:

1. Explicitly discuss generalizability concerns (as related to the question at hand)
2. Explicitly assess relevant generalizability of performance
3. Consider the use of multi-datasets ML techniques

Explicitly consider generalizability concerns as related to the question at hand

In a recent review, Low et al. (2020) emphasize that many studies of voice in psychiatric conditions do not measure out-of-sample performance, e.g., via cross-validation or testing on hold-out subsets of the datasets. Our findings indicate that even when correctly applied these forms of out-of-sample validation of ML models might not provide reliable measures of generalizable performance, depending on the research goals and application needs. For instance, a study trying to identify whether there is an acoustic profile of autism in general should explicitly discuss that this means the findings should generalize across biological sex, language, socio-demographic and ethnic groups, but also that it does not need to generalize across recording settings, given the research question at stake. On the contrary, a study trying to provide tools for the assessment of autism via voice markers across a variety of clinics and contexts should also consider varying recording setups (e.g., microphones, background noise, reverberation differences), but might have lower cross-linguistic generalizability needs (depending on the context of use). Needless to say, requirements will often be domain-specific, but explicitly stating them is a prerequisite for proper usage and cumulative knowledge building within the field, and there has been at least one attempt at establishing such a practice for more traditional statistical approaches (Simons et al., 2017).

Explicitly assess relevant generalizability of performance

Additionally, we argue that generalizability of the algorithm must not only be discussed, but also thoroughly tested and documented. An explicit assessment of generalizability eases the understanding of a model's capabilities as well as its limitations within the intended use. A first step could be an error analysis, as assessing which errors are made by the model might highlight specific biases, e.g. poor ability to identify autistic women, or participants from specific social and ethnic groups (Achenie et al., 2019). However, that is not enough. As we have shown, relying on data from one study and within-study out-of-sample validation is not a reliable way of ensuring generalizable findings, or even documenting limits of the algorithms and data. This, of course, requires the construction or availability of multiple datasets. The increase in collaboration across labs, construction of consortia such as EU-AIMS, ManyBabies, and Psychological Science Accelerator and the increased spread of responsible open science practices might help with this (Bergmann et al., 2016; Murphy & Spooren, 2012).

Consider the use of multi-dataset ML techniques

Finally, as data from multiple studies become available, new techniques become possible, moving from single study training to multi-study approaches, where models are trained on data collected under diverse settings. A basic approach is simply to train the models on multiple datasets, to better account for the heterogeneity between individuals, contexts, tasks and languages. However, more nuanced approaches have been produced providing modular and more flexible ways of taking advantage of multiple datasets. One such approach is the use of mixture-of-experts (MoE) models. In a MoE-model, separate models are trained, each on a different dataset. When evaluating test cases, each model provides both a prediction, and an evaluation of how similar the single case is to the training set (similarity score). The predictions of each expert are then combined into a final prediction, i.e. by weighting the predictions of the experts proportionally to their similarity score (Sechidis et al., 2021). Thus, if one seeks to construct a model that generalizes to three languages, one could train a model on each language separately and combine their predictions in a MoE-model. If a new language were to become available, a new model could be trained only on the new data and added to the mixture. This approach has already shown promising results in related fields, predicting the emotional content of Parkinson's and depressive speech in languages on which the models had not been trained (Hansen et al., 2021; Sechidis et al., 2021). Analogously modular approaches have been developed for language models, including novel techniques to mix, re-weight, add or entirely remove “experts” according to the task at hand (Gururangan et al., 2021). These models can indeed generalize and adapt to new domains and achieve high performance scores when taking advantage of domain-specific expertise. Another promising line of research is transfer learning, where complex models can be trained on large and more easily accessible datasets of non-clinical speech, and then only re-trained on the smaller clinical datasets; or across clinical datasets (Vásquez-Correa et al., 2019). Developing such techniques within the field of voice-based classification of ASD is a highly promising venue to foster higher generalizability.

Finally, we emphasize the importance of sharing openly and fully the modeling process - e.g., strategies to choose hyperparameters, and actual parameter values - to better enable collective investigation of generalizability issues. Accordingly, our choices are fully described in the manuscript and appendices and the code is available on the OSF repository connected to this paper.

Conclusion

This work investigated the generalizability of ML approaches to autistic voices. While promising algorithms could be retrained on new datasets with performance above chance, variability in performance across algorithms trained on different studies was substantial. Further, the models did not generalize well to new similar tasks and not at all to new languages. We argue that greater emphasis must be placed on the generalizability of machine learning models of autistic voices. We recommend to 1) explicitly discuss generalizability concerns (as related to the question at hand), 2) explicitly assess relevant generalizability of performance and 3) consider the use of multi-study ML techniques.

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), 1–16. https://doi.org/10.1016/S0885-2014(00)00014-9

Achenie, L. E. K., Scarpa, A., Factor, R. S., Wang, T., Robins, D. L., & McCrickard, D. S. (2019). A Machine Learning Strategy for Autism Screening in Toddlers. *Journal of Developmental and Behavioral Pediatrics : JDBP*, *40*(5), 369–376. https://doi.org/10.1097/DBP.0000000000000668

Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. (2017). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. *2017 International Conference on Platform Technology and Service (PlatCon)*, 1–5. https://doi.org/10.1109/PlatCon.2017.7883728

Baltaxe, C. A. M., & Simmons, J. Q. (1985). Prosodic Development in Normal and Autistic Children. In E. Schopler & G. B. Mesibov (Eds.), *Communication Problems in Autism* (pp. 95–125). Springer US. https://doi.org/10.1007/978-1-4757-4806-2_7

Benson, V., & Fletcher-Watson, S. (2011). Eye movements in autism. In *Oxford Handbook of Eye Movements* (pp. 709–730). Oxford University Press.

Bergmann, C., Frank, M. C., Gonzalez, N., Bergelson, E., Cristia, A., Ferguson, B., Struhl, M. K., Soderstrom, M., Yurovsky, D., & Byers-Heinlein, K. (2016). *ManyBabies*. https://osf.io/rpw6d/

Bone, D., Chaspari, T., Audhkhasi, K., Gibson, J., Tsiartas, A., Van Segbroeck, M., Li, M., Lee, S., & Narayanan, S. S. (2013). Classifying language-related developmental disorders from speech cues: The promise and the potential confounds. *INTERSPEECH*, 182–186.

Brownlee, J. (2020, April 16). How to Develop Voting Ensembles With Python. *Machine Learning Mastery*. https://machinelearningmastery.com/voting-ensembles-with-python/

Buracas, G. T., & Albright, T. D. (1993). The role of MT neuron receptive field surrounds in computing object shape from velocity fields. *Proceedings of the 6th International Conference on Neural Information Processing Systems*, 969–976.

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. https://doi.org/10.1002/aur.1630

Chekroud, A. (2018). T107. Why Validation Matters: A Demonstration Predicting Antipsychotic Response Using 5 Rcts. *Schizophrenia Bulletin*, *44*(Suppl 1), S157.

Drimalla, H., Scheffer, T., Landwehr, N., Baskow, I., Roepke, S., Behnia, B., & Dziobek, I. (2020). Towards the automatic detection of social biomarkers in autism spectrum disorder: Introducing the simulated interaction task (SIT). *Npj Digital Medicine*, *3*(1), 1–10. https://doi.org/10.1038/s41746-020-0227-5

Eyben, F. (2015). *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Association for Computing Machinery, New York, NY, USA*. https://doi.org/10.1145/1873951.1874246

Fusaroli, R., Grossman, R., Bilenberg, N., Cantio, C., Jepsen, J. R. M., & Weed, E. (2021). 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. *BioRxiv*.

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

Fusaroli, R., Weed, E., Lambrechts, A., Bowler, D., & Gaigg, S. (2018). *Towards a Cumulative Science of Prosody in ASD*. https://www.autism-insar.org/general/custom.asp?page=2018AnnMtg

Grossman, R. B., Edelson, L. R., & Tager, -Flusberg Helen. (2013). Emotional Facial and Vocal Expressions During Story Retelling by Children and Adolescents With High-Functioning Autism. *Journal of Speech, Language, and Hearing Research*, *56*(3), 1035–1044. https://doi.org/10.1044/1092-4388(2012/12-0067)

Gururangan, S., Lewis, M., Holtzman, A., Smith, N. A., & Zettlemoyer, L. (2021). DEMix Layers: Disentangling Domains for Modular Language Modeling. *ArXiv:2108.05036 [Cs]*. http://arxiv.org/abs/2108.05036

Hansen, L., Zhang, Y.-P., Wolf, D., Sechidis, K., Ladegaard, N., & Fusaroli, R. (2021). *A Generalizable Speech Emotion Recognition Model Reveals Depression and Remission* (p. 2021.09.01.458536). https://doi.org/10.1101/2021.09.01.458536

Hegde, S., Shetty, S., Rai, S., & Dodderi, T. (2019). A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, *33*(6), 947. e11-947. e33.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Kirk, M. (2017). *Thoughtful machine learning with Python: A test-driven approach*. O'Reilly Media, Inc.

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

Kuhn, M., Johnson, K., & others. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., & Western Psychological Services (Firm). (2008). *Autism diagnostic observation schedule: ADOS manual*. Western Psychological Services.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96–116. https://doi.org/10.1002/lio2.354

McCann, J., & Peppé, S. (2003). Prosody in autism spectrum disorders: A critical review. *International Journal of Language & Communication Disorders*, *38*(4), 325–350. https://doi.org/10.1080/1368282031000154204

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

Murphy, D., & Spooren, W. (2012). EU-AIMS: A boost to autism research. *Nature Reviews Drug Discovery*, *11*(11), 815–816. https://doi.org/10.1038/nrd3881

Olsen, L. R. (2018). *Automatically diagnosis mental disorders from voice*. Bachelor Thesis presented at Aarhus University.

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. https://doi.org/10.1016/j.schres.2019.11.031

Parola, A., Simonsen, A., Bliksted, V., Zhou, Y., Ubukata, S., Kölkebeck, K., Lund Pedersen, H., & Fusaroli, R. (2020). VOICE PATTERNS IN SCHIZOPHRENIA: A CROSS-LINGUISTIC REPLICATION OF PREVIOUS META-ANALYTIC FINDINGS. *Schizophrenia Bulletin*, *46*(Supplement_1), S230–S230.

Patel, S. P., Kim, J. H., Larson, C. R., & Losh, M. (2019). Mechanisms of voice control related to prosody in autism spectrum disorder and first-degree relatives. *Autism Research*, *12*(8), 1192–1210. https://doi.org/10.1002/aur.2156

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

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

Rocca, R., & Yarkoni, T. (2020). *Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction*. PsyArXiv. https://doi.org/10.31234/osf.io/e437b

Rocca, R., & Yarkoni, T. (2021). Putting Psychology to the Test: Rethinking Model Evaluation Through Benchmarking and Prediction. *Advances in Methods and Practices in Psychological Science*, *4*(3), 25152459211026864. https://doi.org/10.1177/25152459211026864

Scheerer, N. E., Shafai, F., Stevenson, R. A., & Iarocci, G. (2020). Affective prosody perception and the relation to social competence in autistic and typically developing children. *Journal of Abnormal Child Psychology*, *48*(7), 965–975.

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

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-training for Speech Recognition. *ArXiv:1904.05862 [Cs]*. http://arxiv.org/abs/1904.05862

Sechidis, K., Fusaroli, R., Orozco-Arroyave, J. R., Wolf, D., & Zhang, Y.-P. (2021). A machine learning perspective on the emotional content of Parkinsonian speech. *Artificial Intelligence in Medicine*, *115*, 102061. https://doi.org/10.1016/j.artmed.2021.102061

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. https://doi.org/10.1109/MLSP.2019.8918725

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524.

Talkar, T., Williamson, J. R., Hannon, D. J., Rao, H. M., Yuditskaya, S., Claypool, K. T., Sturim, D., Nowinski, L., Saro, H., & Stamm, C. (2020). Assessment of speech and fine motor coordination in children with autism spectrum disorder. *IEEE Access*, *8*, 127535–127545.

Tang, Y., Kleeman, R., & Moore, A. M. (2005). Reliability of ENSO Dynamical Predictions. *Journal of the Atmospheric Sciences*, *62*(6), 1770–1791. https://doi.org/10.1175/JAS3445.1

Trecca, F., Tylén, K., Højen, A., & Christiansen, M. (2021). The puzzle of Danish: Implications for language learning and use. *Language Acquisition*.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual* (Version 3) [Computer software]. CA: CreateSpace.

Vandenbroucke, J. P., Elm, E. von, Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., & Initiative, for the S. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLOS Medicine*, *4*(10), e297. https://doi.org/10.1371/journal.pmed.0040297

Vargason, T., Grivas, G., Hollowood-Jones, K. L., & Hahn, J. (2020). Towards a Multivariate Biomarker-Based Diagnosis of Autism Spectrum Disorder: Review and Discussion of Recent Advancements. *Seminars in Pediatric Neurology*, *34*, 100803.

Varoquaux, G., & Cheplygina, V. (2021). How I failed machine learning in medical imaging—Shortcomings and recommendations. *ArXiv:2103.10292 [Cs, Eess, Stat]*. http://arxiv.org/abs/2103.10292

Vásquez-Correa, J. C., Arias-Vergara, T., Rios-Urrego, C. D., Schuster, M., Rusz, J., Orozco-Arroyave, J. R., & Nöth, E. (2019). Convolutional Neural Networks and a Transfer Learning Strategy to Classify Parkinson's Disease from Speech in Three Different Languages. In I. Nyström, Y. Hernández Heredia, & V. Milián Núñez (Eds.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (pp. 697–706). Springer International Publishing. https://doi.org/10.1007/978-3-030-33904-3_66

Verde, L., De Pietro, G., & Sannino, G. (2018). Voice disorder identification by using machine learning techniques. *IEEE Access*, *6*, 16246–16255.

Williamson, J. R., Quatieri, T. F., & Smith, K. M. (2017). *Vocal Markers of Motor, Cognitive, and Depressive Symptoms in Parkinson's Disease*. MIT Lincoln Laboratory Lexington United States.

Yankowitz, L. D., Schultz, R. T., & Parish-Morris, J. (2019). Pre- and Paralinguistic Vocal Production in ASD: Birth Through School Age. *Current Psychiatry Reports*, *21*(12), 126. https://doi.org/10.1007/s11920-019-1113-1

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685