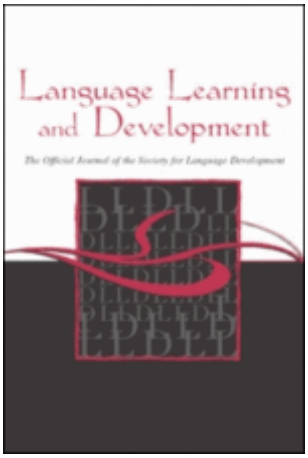


Different in different ways: A network-analysis approach to voice and prosody in Autism Spectrum Disorder

Journal:	*Language Learning and Development*
Manuscript ID	HLLD-2022-0074.R1
Manuscript Type:	Research Article
Date Submitted by the Author:	03-Jan-2023
Complete List of Authors:	Weed, Ethan; Aarhus University, Fusaroli, Riccardo; Aarhus University Simmons, Elizabeth; Sacred Heart University, Communication Disorders Eigsti, Inge-Marie; University of Connecticut, Psychological Sciences;
Keywords:	Speech production, Computational, Pragmatics

Introduction

Difficulties with speech and language are a defining feature of autism spectrum disorder (ASD) (Association, 2013), and poor communication skills exacerbate the risk being bullied five-fold (Cappadocia, Weiss, & Pepler, 2012). Atypical prosody and vocal presentation impact social integration and employment opportunities. Indeed, some older research suggests that unusual prosody is among the first features that elicit an impression of oddness from others (Mesibov, 1992; Shriberg & Widder, 1990; Van Bourgondien & Woods, 1992). Even individuals with ASD whose expressive and receptive language scores are in the typical range have difficulty with prosody (Shriberg, Paul, Black, & Santen, 2011), and children with ASD are rated as more socially awkward on the basis of audio speech samples alone (Grossman, 2015; Redford, Kapatsinski, & Cornell-Fabiano, 2018; Sasson et al., 2017). One mechanism suggested for this is that speakers with ASD are more consistent in their phonetic production than NT speakers, as reported in a study of adults with and without ASD (Kissine, Geelhand, Philippart, Harmegnies & Deliens, 2021). Atypical vocal quality is one of the signals of the "frank" autism presentation, apparent to expert clinicians within just a few seconds (Marchena & Miller, 2017). However, despite their significant clinical impact, the specific acoustic features underlying speech differences in ASD are as yet poorly understood. This is a clinically important gap in knowledge, as a better understanding of what acoustic cues contribute to atypical speech quality would enable us to develop better-informed targets for intervention. The goal of this manuscript is to combine information from acoustic analyses with ratings from naïve and clinical raters, in order to provide some insights into what acoustic qualities make a voice distinctively autistic.

Although clinicians since Asperger have described the voices of people with autism as having unusual prosodic (e.g., "robotic", "flat", "monotone"), and vocal (e.g., "harsh", "nasal"

and "hoarse") qualities, there is little consensus on exactly *which* acoustic features of ASD speech differ from typical speech (McCann & Peppé, 2003). This is true even though productions may be reliably reported as "odd" by naïve listeners, as was the case for a study of utterances produced by 12 youths with Asperger syndrome (Filipe, Castro & Vicente, 1981). A recent systematic review and meta-analysis (Fusaroli, Lambrechts, Bang, Bowler, & Gaigg, 2017) highlights several potential acoustic contributors to group differences. The meta-analytic findings point to higher pitch, and increased pitch variability, number of pauses, and pause duration, although replication attempts show these patterns might not be generalizable across languages and samples (Parola, 2022; Fusaroli et al, 2021, Rybner et al, 2022). While some studies have found that acoustic differences map onto expert clinician ratings of autism-specific symptoms (McCann et al, 2007; Study 1, Diehl et al. 2009), these findings have not always replicated, even within the same lab (Study 2, ibid; Fusaroli et al 2021).

Although the speech of autistic people is usually described as having an atypical *prosody*, the adjectives used to describe the speech often allude to aspects that go beyond prosody. In one of the earliest descriptions of the voices of children with autism, for instance, Asperger (1991) described them as "shrill," "soft," and "nasal," qualities that belong more properly to the realm of *voice quality*. Voice quality has been shown to play an important role in the impressions we form of speakers. Listeners routinely extract numerous impressions of speaker characteristics, based on voice quality alone: gender identity, age, mood, socio-economic status, sexual orientation, social relationships, and even neuropsychiatric conditions (Bryant et al., 2016; Cummins et al., 2015; Parola, Simonsen, Bliksted, & Fusaroli, 2020; Low, Bentley & Ghosh, 2020, Redford et al., 2018; Weed & Fusaroli, 2020). However, a one-to-one mapping between acoustic features and perceptual impressions has proven elusive; this likely reflects broader

challenges in mapping acoustic features onto other dimensions of speech, such as phonology, described as the "many to many" mapping problem (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

Both expert and non-expert raters are able to reliably distinguish between typical and atypical prosody in short samples of speech, and are thereby often able to predict which speakers have an ASD diagnosis (Nadig and Shaw, 2012; Redford et al., 2018). However, consistent with the complexity noted above, listeners, including trained clinicians, are often not explicitly aware of which acoustic characteristics seem distinctive in ASD, and thus of which acoustic features or combination of features they are responding to. Nadig and Shaw (2012) found that speech-language pathology students could not reliably identify atypical variation in pitch in speakers with ASD, although the same raters were able to distinguish between broadly typical versus atypical prosody in the same speakers. In a related study, Redford et al. (2018) found that naïve raters were able to distinguish between the voices of autistic and neurotypical (NT) children when asked to score on atypicality, but not when listening to low-pass or high-pass filtered versions of the same speech, suggesting that "[…] sound pattern differences are subtle enough to be obscured in degraded speech," (Redford et al., 2018, p. 289). Dahlgren et al. (2018) found that a group of three speech language pathologists who specialized in voice were only able to correctly identify three out of eleven children with ASD on the basis of voice alone, and concluded that voice and speech characteristics are not unproblematic predictors of diagnosis.

To further complicate matters, there is no consensus as to what is distinctive about the voice and prosody of people with ASD. Peppé et al. (2007) point to the wide range of sometimes contradictory adjectives that have been used to describe the speech of people with ASD. They note that autistic speech has been simultaneously described as "dull", "wooden", "singsong",

"robotic", "stilted", "over precise" and "bizarre", but also both "monotonous" and "exaggerated" as well as "slow" and "fast," raising the question of whether it possible to draw any consistent conclusions about expressive prosody in ASD at all.

There are several possible explanations for these inconsistent findings. First, it may be the case that there simply are no consistent patterns to detect. Another possible explanation is that there are different *profiles* of atypical prosody and voice quality in speakers with ASD. In other words, the atypical prosody and voice quality of speakers with ASD may fall into a small (or large) set of distinct patterns, with several different characteristic types of atypicality. This second explanation is consistent with both the reliable differences in perceptual impressions, and with the lack of a single clear constellation of acoustic features characteristic of the speech of people with ASD. While listeners may pick up on these atypicalities, the presence of multiple different constellations of acoustic features that are all "atypical" may make it difficult to reliably predict diagnosis on the basis of acoustic features alone. As McCann and Peppé write, "If findings were consistent, small-scale studies would offer pointers, but as it is these do not inspire confidence," (2003, p. 347). This opinion is reiterated a decade later in Fusaroli et al. (2017).

Prosody includes changes in pitch, loudness, and rhythm of sounds (Cutler et al., 1997; Peppé, 2009; Shattuck-Hufnagel & Turk, 1996). *Pitch* describes the perceptual experience of the fundamental frequency (F0) of a sound (e.g., a high versus low voice), and *variation in pitch* is most often described by the standard deviation, range, or interquartile range of F0. *Volume* or loudness refers to the perceptual experience of intensity, or the amount of energy in the sound waves. Finally, the concept of *rhythm* describes acoustic features related to the duration of speech and pauses: among them the duration of the utterance, the number and length of pauses, as well as speech rate and articulation rate. While a high speech rate relates to the amount of

information (syllables) per unit of time, a high articulation rate indicates more clipped, less drawn-out syllables. An individual could thus have, for example, a low speech rate but a high articulation rate; such a speaker would produce very quick syllables, but have long pauses in their speech. All these measures have been indicated as potential markers of ASD (Fusaroli et al., 2017).

Although descriptions of voice quality appear in the literature (e.g., "harsh", "nasal", or "hoarse"), voice quality measures are remarkably absent from the study of vocal atypicality in ASD. Two studies (Bone et al., 2014; Kissine & Geelhand, 2019) have investigated *jitter*, that is, cycle-to-cycle changes in the fundamental period, which is associated with the perceptual qualities of breathiness and hoarseness (Eskenazi, Childers, & Hicks, 1990; Wolfe, Fitch, & Martin, 1997). Other candidate acoustic features are *shimmer* (Kissine & Geelhand, 2019), which quantifies cycle-to-cycle fluctuations in the amplitude of the waveform and has been related to both breathiness and hoarseness (Wolfe et al., 1997); *harmonics-to-noise ratio*, which quantifies the relative amount of energy in harmonic portions of the spectrum with other, "noise" energy, and has been related to hoarseness (Yumoto, Gould, & Baer, 1982); and *H1-H2*, the relative amplitudes of the first two harmonics, which is also linked to the perception of breathiness (Hillenbrand & Houde, 1994; Klatt & Klatt, 1990; Kreiman & Gerratt, 2010). Not only do all these features measure voice qualities which have been linked to perceptual qualities noted in ASD, they also tend to be intercorrelated, as they all derive from fluctuations in the glottal source (Murphy, 2000).

We have discussed a handful of the most common features, but many more can be extracted. In studies on dysarthric speech, Borrie et al. (2020) measured over 800 acoustic features, and Al-Qatab & Mustafa (2021) measured over 5000 acoustic features. The ComParE

baseline feature set consists of over 6000 features (Schuller et al., 2016). With so many potentially important features, the task of feature selection is critical. Unfortunately, although several different methods are commonly used, there is currently no single accepted method for reducing the number of features to include in an analysis. Often, features are chosen with the aim of finding the best combination of features for predicting category membership or some other measure (e.g., predicting diagnosis, or symptom severity). Borrie et al (2020) used independent component analysis (ICA) for feature reduction. ICA seeks to identify independent "sources" of information in a multivariate dataset. Al-Qatab et al (2021) employed seven different algorithmic feature reduction techniques: Interaction Capping, Conditional Information Feature Extraction, Conditional Mutual Information Maximization, Double Input Symmetrical Relevance, Joint Mutual Information, Conditional Redundancy, and Relief. Another common method is Principal Component Analysis (PCA; e.g. Mittal & Sharma, 2021; Peng et al., 2007). While these algorithmic methods can be a powerful way to reduce the number of features to a manageable number, they have the disadvantage of often resulting in feature sets that can be difficult to interpret in ways that are clinically useful. PCA, for example, reduces individual features to linear combinations of features, which can be difficult to describe intuitively.

Given these challenges, the present study adopted an alternative approach. We chose a small number of acoustic features that were among the most frequently mentioned in the literature: a measure of pitch variation (standard deviation of fundamental frequency); two measures of rhythm (speech rate; syllables per second) and articulation rate (syllables per second after removing pauses, which were defined as absence of voice for more than 200 milliseconds); and a measure of voice quality (jitter). Note that although variation in fundamental frequency and jitter are both derived from the glottal source, they are independent: fundamental frequency

(perceived as pitch) is the lowest of the many harmonics in the voice, while jitter represents the difference in length between each cycle and the preceding cycle in the sonogram; put broadly, pitch is estimated by looking for large-scale similarities in the acoustic signal, while jitter is estimated by measuring small differences in the signal. As measures of rhythm, speech rate and articulation rate are complementary, capturing two potential sources of the "robotic" quality that is often ascribed to autistic speech. Furthermore, given that elicited sentences were scripted, and the number of syllables was therefore held constant, combining speech rate and articulation rate also provided an implicit measure of pause length. We chose not to include intensity, although it is among the key elements of prosody, as our participants were at variable distances from the microphone during recording, which influences intensity. Although algorithmic feature reduction might well result in a set of features with a higher predictive power, our goal in choosing common features that have been previously investigated in speakers with ASD, and which cover, to the extent possible, qualitative reports on the speech of people with ASD, was to select a set of easily-interpretable features, which would be relevant to clinicians.

Certainly, our choice of features presents some disadvantages. Standard deviation of F0 is a very crude measure of pitch as a contributor to prosody and masks many important nuances of pitch modulation. Jitter can be affected by factors such as vocal strain (Brockmann-Bauser et al., 2014; Huang et al., 1995), although jitter may be less affected by room noise and microphone quality than other measures such as shimmer and harmonics-to-noise ratio (Bottalico et al., 2020; van der Woerd et al., 2020). In this exploratory study, however, the advantage of working with a small number of commonly-used features outweighed the disadvantages.

Rather than trying to identify a single set of acoustic characteristics that describe the "autistic voice," the current study aims to identify acoustic profiles that characterize clusters of

individuals whose voice and prosody are more similar to each other than to speakers in a different cluster, and to test how these profiles align to diagnosis and to listener ratings. We posed the following exploratory hypotheses:

H1. Modelling speakers as nodes of acoustic features in a network will allow us to identify coherent clusters of speakers.

H2. Neurotypical and autistic speakers will tend to cluster differently.

H3. Clusters of speakers will be characterized by distinctive constellations of acoustic features.

H4. Clusters will reflect the subjective ratings of naïve and clinical raters.

Network analysis techniques are well suited to this sort of analysis, because they allow visualization of individuals in relation to each other. To address these hypotheses, we coded speech samples, and used techniques from network analysis to build clusters of speakers, derived from the acoustic features of their speech. We then inspected the distribution of autistic and neurotypical speakers across the clusters, as well as the acoustic profiles that generated these clusters. Finally, we tested whether naïve and expert raters tended to rate speakers from the clusters as more or less atypical. We emphasize that our approach here is exploratory, and acknowledge that our sample size is limited. To our knowledge, the methods we employ here for identifying speaker clusters on the basis of acoustic measures have not previously been used for this purpose, although similar graph-based methods are currently being explored in the development of speaker recognition software (Chen et al., 2021; Wang et al., 2021).

Different in different ways

Methods

Participants

Participants were 13 autistic and 13 neurotypical (NT) male adolescents, all native speakers of American English. Full-scale IQ (FSIQ), verbal IQ (VIQ), and non-verbal IQ (NVIQ) scores were estimated with the Stanford-Binet Abbreviated IQ (Roid, 2003). Groups did not differ in chronological age, full-scale IQ, or structural language abilities, measured using the Clinical Evaluation of Language Fundamentals (CELF; Wiig, Semel, & Secord, 2003). CELF and IQ scores were in the average range for all participants. ASD diagnoses were confirmed by trained graduate clinicians using the Autism Diagnostic Observation Schedule (ADOS) Module 3 (Lord et al., 2000), selected sections of the Autism Diagnostic Interview, Revised (ADI- R; Lord, Rutter, & LeCouteur, 1994), and clinical judgment, based on the DSM-IV-TR criteria checklist (APA, 2000). Parents reported no co-morbid diagnoses, and completed the Social Responsiveness Scale (Constantino et al., 2003) to verify group membership of both ASD and NT participants, and as measure of symptom severity. Participant details are shown in Table 1.

Table 1: Participant Characteristics

	ASD (n = 13)	NT (n = 13)	Welch Two Sample t-test
Age (years)	14.24 (1.82)	14.30 (1.41)	t(22.60) = -0.08, p = 9.36
FSIQ Standard Score	102.76 (9.75)	104.15 (9.60)	t(23.99) = -0.36, p = 0.71

Different in different ways

VIQ Standard Score	9.69 (2.56)	11.53 (2.29)	t(23.71) = -1.93, p = 0.06
NVIQ Standard Score	11.30 (2.59)	9.84 (1.95)	t(22.28) = 1.62, p = 0.11
CELF Core Language Standard Score	109.41 (10.42)	114.84 (10.08)	t (22.69) = -1.32, p = 0.19
Social Responsiveness Scale	74.66 (11.52)	44.84 (7.95)	t(19.37) = 7.47, p = 0.00004
ADOS	9.90 (2.84)	N/A	N/A

Speech elicitation

The data were originally collected as part of a study investigating how adolescents with and without ASD used prosody to disambiguate sentences (Mayo, 2015). Although the original study included a training session, in which participants were provided with suggestions on how they could modify their prosody to better disambiguate sentences, the utterances used in the present study were from the pre-training, baseline condition, in which participants each produced eight sentences with similar sentence structures and high-frequency words. The sentences were written on printed cards and participants were asked to learn the sentence; once it was memorized, they were asked to speak the sentence aloud "in a normal voice" without reading from the card. Sentences were all instructions to interact with an animal, e.g., "Point at the lamb with the flower." or "Tap the duck with the lollipop." Speech was recorded with a Marantz Professional Model PMD660 audio recorder. A full description of the elicitation procedure, as well as additional information on recruitment, can be found in Mayo (2015). Although this

method of elicitation lacks the ecological validity of, e.g., recordings of natural conversations, the controlled nature of task was well-suited to the purpose of the current study. Because the speech samples included all identical words, potential prosodic differences due to words with differing numbers of syllables, or pitch fluctuations due to stress patterns or emotional valence associated with different phrases were controlled for, providing a more appropriate framework for our exploratory network analysis than recordings of spontaneous speech.

Feature Extraction

F0, jitter, speech rate, and articulation rate were extracted using custom Praat (Boersma & Weenink, 2001) scripts (all scripts and extracted data are available on OSF at https://osf.io/zw67n/?view_only=63441e15ec264184bb7b0b22c4140a22). For F0 extraction, we set a floor value of 50 Hz, and a ceiling of 400 Hz. We inspected individual pitch tracks for evidence of multimodality (period doubling). Fig S1 (available at the OSF repository) displays histograms of the results of the pitch tracking for all participants. Although a few of the speakers voices did show some evidence of multimodality (e.g., participants 9027, 9030, 9033), on closer inspection at least some of these, such as participants 9027 and 9033, could be accounted for by a period of very low, occasionally "creaky" phonation. For the sake of consistency, we applied the 50-400 Hz range to all speakers. F0 was converted to a semitone scale, using the hqmisc (Quene, 2022) package for R, using the default value of 50 Hz as the reference frequency.

Network Models and Community Detection

To identify groups of speakers with similar acoustic patterns, we first constructed a network of speakers. We used the *qgraph* (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012) package for R Statistical Software (v4.2.1; R Core Team, 2022) to create a partial correlation matrix of speakers, with each speaker represented by the four selected prosody

and voice features. We plotted this matrix as a network graph, in which each speaker is represented by a circle (node); the lines (edges) connecting the nodes represent the pairwise correlations (similarity) between each pair of speakers, when the effects of all other speakers have been removed; see Figure 1.

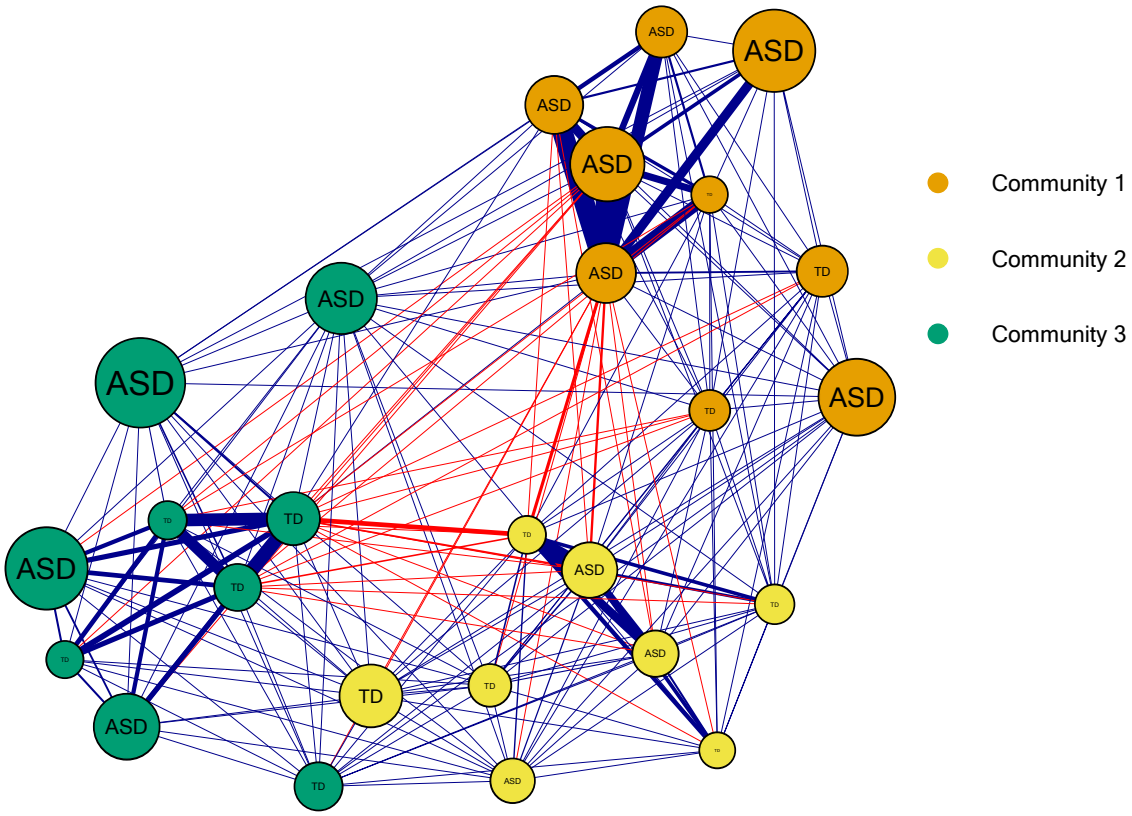


Figure 1: Network of acoustically-defined communities. Each node represents a single speaker. Blue connections between nodes represent positive correlations, red lines represent negative correlations. Line thickness indicates the absolute strength of the correlation.

We used the edges between nodes to identify clusters or "communities" of speakers with similar voice/prosody profiles. To do this, we used a spin glass community detection algorithm,

as proposed in Reichardt & Bornholdt (2006) and Traag & Bruggeman (2008), and implemented in the *iGraph* package (Csardi & Nepusz, 2006) for R. The spin glass approach attempts to find communities of nodes by maximizing the number of *positive edges* (positive correlations) between nodes within the community and minimizing the number of positive edges between members of the community and members outside the community. At the same time, the algorithm maximizes the number of *negative edges* (negative correlations) with nodes outside the community, while minimizing the number of negative edges within the community. Although the user sets an upper limit to the number of communities that the algorithm will detect, the optimal number of communities may be lower than this upper limit (Csardi & Nepusz, 2006). We set the upper limit of possible communities at ten. Because the algorithm is non-deterministic, it may not always settle on the same number of communities each time it is run. For this reason, we followed the procedure outlined in Djelantik et al. (2020) and seeded the random number generator with a seed which would obtain the median number of communities identified in 1000 runs of the algorithm. Using this method, the algorithm always settled on a three-community solution, and we therefore chose a seed which produced a 3-community solution, and used these three communities for further analysis.

Listener judgments

Ten clinicians with ASD expertise and 15 undergraduates without ASD experience, all naïve to the study hypotheses, listened to unaltered speech samples, and rated each sample as "atypical or unusual" on a 1-3 scale (1 = typical; 2 = somewhat unusual; 3 = definitely atypical; undergraduates) or as autistic or non-autistic (0 = NT-like, 1 = ASD-like) on a binary scale (clinicians). The clinical and naïve raters' scores were calculated as the mean of all the raters' scores for speech samples from that individual, normalized to a 0 to 1 scale. Interrater reliability

was calculated as Average Score Intraclass Correlation, using the *irr* package (Gamer et al., 2012) for R. Estimates were based on a mean rating (naïve: k = 17, expert: k = 5) agreement, 2-way random-effects model (Koo & Li, 2016).

Results

Community Detection. Community membership was very consistent. Out of 1000 runs, there were five speakers who were occasionally placed in different communities; see Table 2. The most inconsistently-grouped was an NT speaker, who was placed in Community 1 in 67.9% of runs and in Community 2 in 32.1% of runs. Another NT speaker was placed in Community 3 in 70.3% runs and in Community 2 in 29.7% runs. The other three were placed in other communities in less than 2% of runs. Each of these speakers were placed in their "preferred" community in the final network.

Table 2: Community assignments across 1000 runs (inconsistently assigned participants only). All participants not listed in the table were assigned to the same community across all runs. Values indicate percentage of time the participant was assigned to the community.

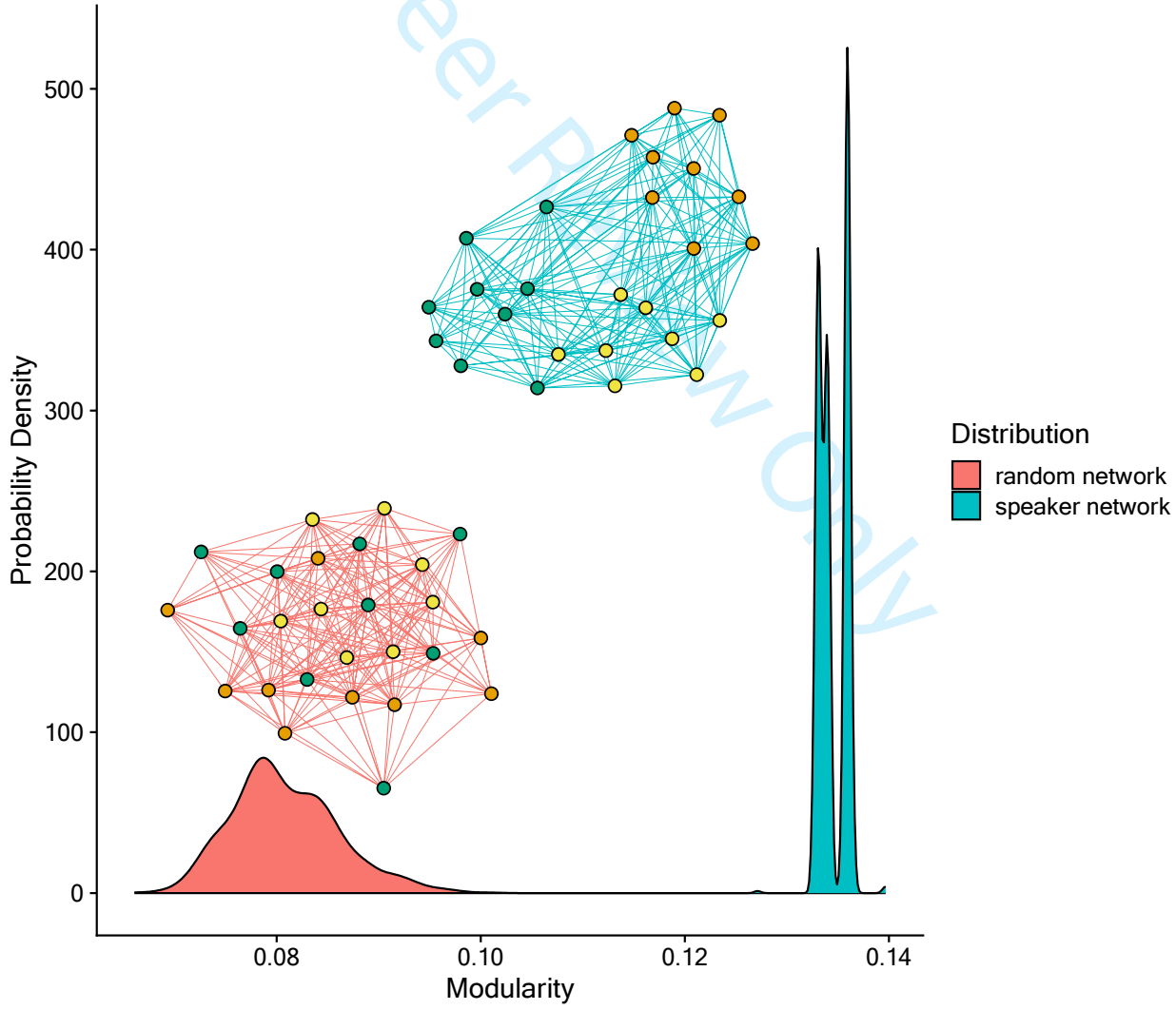
ID	Dx	Com1	Com2	Com3
9004	NT	67.9	32.1	0
9009	ASD	1.2	0	98.8
9021	ASD	99.9	0.1	0
9025	NT	0	98.6	1.4
9029	NT	0	29.7	70.3

Different in different ways

Network modularity. Because the community-detection algorithm can identify "communities" even in randomly-generated networks, it is important to assess the strength of the communities detected. This can be done by calculating a "modularity" score, that is, an estimate of the degree of "clumpiness" in the network, and then comparing this score with the distribution of modularity scores from randomly-generated networks. To test whether the placement of speakers into communities in our network was at greater than chance level, we first computed the level of modularity of the network, where modularity is defined as the number of edges that fall within a community divided by the number of edges that would be expected to fall within a community in a randomized network, in which each node has the same number of connections as the actual network, but in which these connections are randomized (Clauset et al, 2004). This resulted in a modularity score of 0.1345. We then compared the modularity of our network with the distribution of modularity scores from a population of random networks built with the same constraints of numbers of nodes and number of edges incident upon each node (Csárdi et al, 2016; Maslov & Sneppen, 2002) by generating 1000 random networks with the same constraints as our network of speakers, detecting communities for each of them using the same spin glass algorithm, and calculating the modularity score for each random network. The resulting mean modularity score was 0.0808, with a standard deviation of 0.0052. Thus, the modularity score of our network lay approximately eleven standard deviations above the mean; see Figure 2. By Chebyshev's inequality (Knuth, 1997 pg. 98), no more than 0.82 percent of values can lie 11 standard deviations away from the mean of the probability distribution of modularities calculated on random networks, implying that it is highly unlikely that our network was drawn from the same distribution as the random networks. To further quantify the modularity of the speaker network, we compared the distribution of modularities from 1000 runs of the community

detection algorithm on the actual data with the modularities from 1000 random networks with the same constraints as the speaker network with a Welch Two Sample t-test, $t(1315.8) = -337.42$, $p = 2.2^{-16}$)

Clustering of Diagnosis and Acoustic Features. Community 1 (Fig. 1, orange group; $n = 9$) consisted of six autistic speakers and three NT speakers, Community 2 (Fig. 1, yellow group; $n = 8$) consisted of five NT speakers and three autistic speakers, and Community 3 (Fig. 1, green group; $n = 9$) consisted of four autistic speakers and five NT speakers.



Different in different ways

Figure 2: Distributions of modularity scores for 1000 random networks built with the same constraints as the actual network of speakers, and the distribution of modularity scores for 1000 runs of community detection on the speaker network. Also displayed are a sample random network and the actual speaker network. The colors of the connections in the sample networks indicate which distribution they are drawn from. The colors of the nodes indicate community membership.

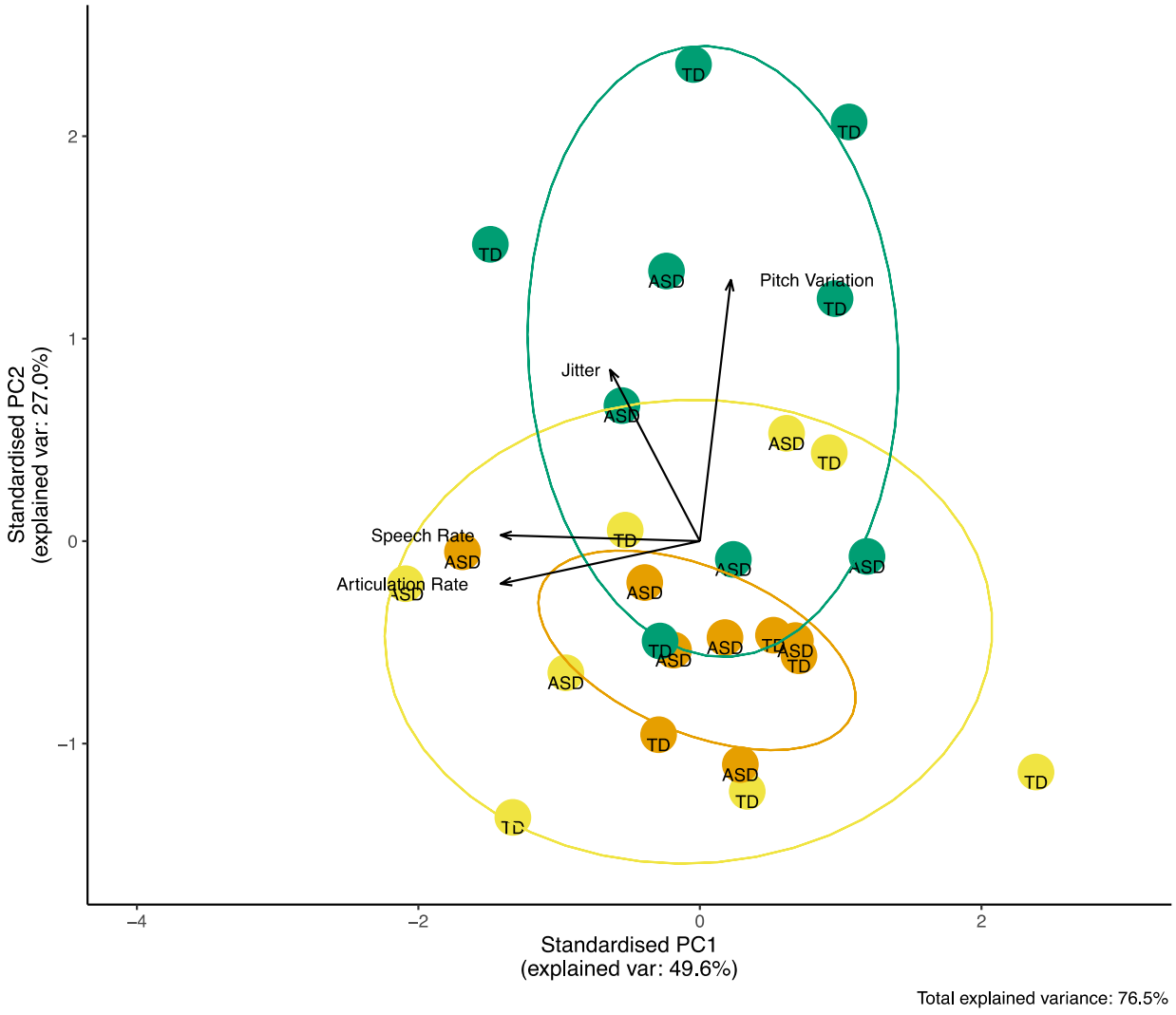


Figure 3: Speakers and communities visualized along the first two dimensions of a principal components decomposition. Ellipsis colors and point colors indicate communities detected by the spin glass algorithm (orange = Community 1, yellow = Community 2, green = Community 3). Arrows indicate the influence of each of the features on the components.

To better visualize the patterns of acoustic variables that characterized the three communities, we plotted these variables in a principal components analysis, using the *pca* function from the AMR package for R (Berends et al., 2021). The first component accounted for 49.6 percent of the variance, and was heavily influenced by the two rhythm features, while the second component accounted for 27 percent of the variance, and was dominated by pitch variation and jitter, both of which are derived from the glottal source. Communities 1 and 2 (orange and yellow) occupied a similar area in the PCA space, although speakers in Community 2 covered a broader range than speakers in Community 1. Community 3 (green), on the other hand, was more heavily influenced by the glottal measures, in particular pitch variation.

Subjective ratings. Clinical and naïve ratings were very highly correlated, $r(24) = 0.84$, $p < 0.0001$, with high interrater reliability (intraclass correlation coefficient): naïve raters: ICC = .957, $F(26,192) = 29.9$, $p < .0001$, CI(95%) $0.929 < ICC < 0.978$; expert raters: ICC = . 882, $F(26,60.6) = 10.1$, $p < .0001$, CI(95%) $= 0.785 < ICC < 0.941$. Separate 2-way ANOVAs for naïve and expert raters showed a significant main effect of diagnosis: naïve: $F(1,22) = 22.18$, $p < .0001$; expert: $F(1,22) = 51.38$, $p < .0001$, indicating that both sets of raters were more likely to rate a talker as atypical or autistic, if that talker was from the ASD group. For the naïve raters there was a significant main effect of Community, but not for the expert: naïve: $F(2,22) = 4.012$, $p < 0.05$, expert: $F(2,22) = 2.070$, $p < .14$). The greatest divergence between naïve and expert raters was in their assessment of the three autistic speakers and one NT speaker who were placed in Community 2. Two of these three autistic speakers were rated at a similar level as the NT speakers by the naïve raters, and a single NT speaker from this community was rated similarly to the autistic speakers from the other communities. The expert raters, on the other hand, while

rating the speakers from Community 2 as slightly less likely to have an autism diagnosis than those in Communities 1 and 3, were much more accurate in their assessment.

Symptom severity. Scores on the Social Responsiveness Scale differed as a function of group (Table 1). However, separate ANOVAs for NT and ASD speakers showed no relation between symptom severity (SRS) and community (NT: $F(1,12) = 0.553$, $p = 0.471$; ASD: $F(1,10) = 0.004$, $p = 0.952$). SRS data were missing for one ASD speaker.

Verbal IQ. Although there was no significant difference between the NT and ASD groups on VIQ (Table 1), because the p-value for this comparison was close to the 0.05 alpha level (0.06), we also performed separate ANOVAs to see whether VIQ variably predicted community membership for these two groups. VIQ was not a significant predictor of community membership for either the NT ($F(1, 12 = 0.01$, $p = 0.917$) or the ASD ($F(1, 10) = 0.502$, $p = 0.495$) groups.

Different in different ways

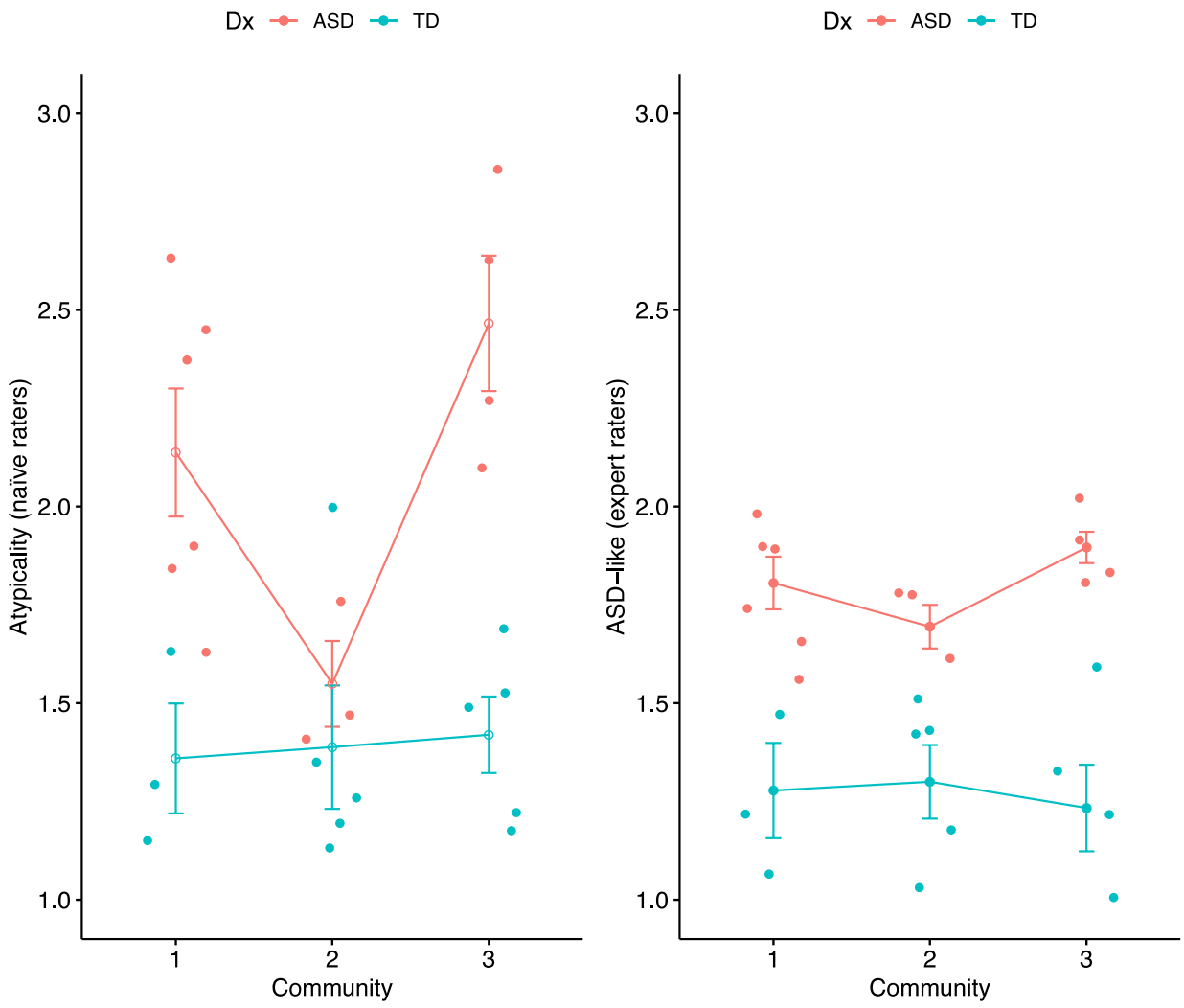


Figure 4: Relationship between ratings by naïve (left) and clinically-trained (right) raters and community membership for ASD (red) and NT (blue) participants

Discussion

This exploratory study aimed to characterize the acoustic features that underlie perceptually distinctive qualitites of autistic speech. Drawing on the scant prior literature examining these features, we expected that variation in fundamental frequency, jitter, speech rate, and articulation rate were likely to be important in differentiating the speech of individuals

with ASD. We compared the acoustic qualities of age- and NVIQ-matched groups of speakers with a history of autism or typical development as they spoke a series of eight identical utterances. Given the failure of prior studies to find generalizable group-level differences in individual acoustic cues, we hypothesized that a network approach might be more successful, given that it permits the assessment of a constellation of features. In contrast to supervised predictive approaches, which attempt to build models of the voice and prosody of individuals with ASD for the purpose of making diagnostic predictions, our approach was exploratory and unsupervised, and thereby made no *a priori* assumptions about the role of diagnosis. This method opens the possibility that multiple prosodic and voice quality profiles can be identified in a data-driven fashion, with the hope that the identification of these profiles can inform not only clinical practice but also future predictive modelling attempts.

Based on the literature, we proposed four hypotheses relevant to identifying a prosodic profile for speakers with ASD.

H1. Modelling speakers as nodes of acoustic features in a network will allow us to identify coherent clusters of speakers.

To our knowledge, this approach of modelling a group of speakers as nodes in a network defined by the similarity or dissimilarity of acoustic properties of the voice has not been previously reported, so our results are necessarily both tentative and exploratory. At the same time, although the number of participants in the study was small, the separation of speakers into groups was quite robust. Given four distinct acoustic measures of prosody and voice as measured in a set of eight utterances (identical across speakers), the algorithm reliably detected three groups of speakers over many runs with different random seeds. Although this method needs further exploration and validation, including experimenting with the inclusion of a wider variety

of acoustic features, and different types of utterances, the current results suggest that it is possible to reliably identify clusters of individual speakers on the basis of acoustic variables.

H2. NT speakers and speakers with ASD will tend to cluster differently.

The speakers were divided into three clusters; ASD and NT speakers were not evenly distributed over the three groups. Approximately 69% of the speakers in Community 1 but only 37.5% of the speakers in Community 2 were autistic. This result suggests that the acoustic measures used by the community-detection algorithm to cluster speakers did to some degree correlate with the presence or absence of ASD. However, Community 3 adds complexity to this picture. While 17 of the speakers were grouped into two communities which might otherwise be called the "autistic" community and the "NT" community, over a third of the participants were clustered in Community 3, which was composed of a nearly equal number of NT and ASD speakers. Clearly, sorting the speakers by diagnostic group cannot rely on acoustic features alone, at least not using the features that we chose. At the same time, both naïve and expert raters were quite successful in their categorization of these speakers.

H3. Clusters of speakers will be characterized by distinctive constellations of acoustic features.

Although our results are preliminary, the combined patterns of acoustic features did form qualitatively different clusters of speech samples. Variation in Communites 1 and 2 were accounted for primarily by differences in the rhythmic features speech rate and articulation rate, while Community 3 was mostly characterized by variation in F0 and jitter.

H4. Clusters will reflect the subjective ratings of naïve and clinical raters.

Both naive and expert raters were very successful in distinguishing autistic and neurotypical speakers. Interestingly, this was the case even for speakers from Community 3, which consisted

of a nearly even mixture of ASD and NT speakers whose voice and speech patterns closely resembled each other, at least as measured with our four acoustic variables. Since the content of the sentences uttered by the speakers was the same, these raters likely relied on subtle characteristics of either intonation, voice quality, or both, which at least partially eluded our algorithmic analysis.

In addition to the limited sample size, there are at least three reasons for the failure of the algorithmic approach to group the speakers into distinct diagnostic group clusters. **First**, the network may have included too few acoustic features, or included some features which are not distinct across groups. **A second possibility** is that the network used the "correct" set of features, but these features were not accurately or appropriately measured or processed. **Finally**, of course, the common perception that there is an "autistic voice" may be inaccurate; perhaps there is in fact no single acoustically-definable profile. We discuss each possibility in turn.

Regarding the first possibility, it is certainly plausible that adding other acoustic features such as shimmer, H1-H2, MFCC's (Mel Frequency Cepstral Coefficients), or raw spectrograms to the mix might improve our ability to algorithmically cluster speakers into groups of primarily ASD or NT speakers. However, this raises new challenges. First, adding more features increases the risk of overfitting. This could be alleviated by a subsequent feature-reduction step, such as using principal components analysis (PCA) to identify composite variables with the greatest predictive power. In addition to overfitting, adding additional features can further contribute to difficulties in interpreting the communities. The most useful outcome for training new clinicians, for example, would be to identify a set of easily-identified acoustic features that have been shown to be characteristic of the speech of autistic people. Neither a long list of relatively complex acoustic features nor an inscrutable composite component from a PCA achieves this

goal (see supporting information S4 of Fusaroli et al (2021) for a discussion of the interpretation of PCA components in acoustic analysis). While likely incomplete, a list of approximately three to five intuitively understandable features may be of the greatest utility for intervention. The prior literature, together with the current results, indicates that pitch variation and rhythm features are likely to be important.

Regarding the second possibility, one could retain the current list of features, with some small adjustments. Pause behavior (as indirectly measured by relationship of speech rate to articulation rate in the present study) appears to be an important distinguishing factor in our data, but the window for defining a pause could be adjusted in either direction; in this study, it was defined as silences of 0.2 seconds or longer. Further, a simple measure like speech rate or articulation rate may not be refined enough to capture the essence of the atypical speech rhythms that our naïve and expert raters picked up on. For example, there is some research suggesting that analysis tools, such as recurrence quantification analysis (RQA), that embrace the temporal aspect of speech and identify recurring patterns over time, can be combined with features such as pause length to identify speakers with ASD (Fusaroli, Bang, & Weed, 2013; Fusaroli, Grossman, Cantio, Bilenberg, & Weed, 2015). However, like feature reduction with PCA, this approach results in a set of predictive features that are less easy to interpret intuitively.

This brings us to the third possibility: that there is no unique or defining "autistic voice." This argument is highly appealing on several grounds. First, the empirical, clinical experience and especially self-reports from autistic people, all emphasize the tremendous heterogeneity of ASD; see Mottron & Bzdok (2020) and Waterhouse (2013) for discussion. As an example, although reduced pitch variation may be typical of many, though not all, autistic people, the speaker in our sample most consistently identified by both naïve and expert raters as either

"atypical" or "likely to be diagnosed with ASD" had a relatively *high* degree of pitch variation. Clearly, an algorithm that builds predictions based alone or in part on pitch variation would be likely to misclassify this speaker as NT, but the human raters were not in doubt. A qualitative assessment of this speaker's speech offers some clues to this apparent contradiction, as discussed further below.

Different in different ways?

The network community-detection algorithm split the speakers into clusters that were related to diagnosis, but did not separate them evenly into two clusters defined by diagnosis. This lends support to the idea that while acoustic features of prosody and voice are related to diagnosis, there is no single acoustic feature or even constellation of features which is consistently associated with ASD or with the perception of atypicality. This tracks with the initial observation that radically different qualitative descriptions have been used of the prosody of people with autism: monotone versus singsong, slow versus fast. There may be a similar underlying cause that results in these seemingly opposite outcomes; for example, an awareness that NT speakers use pitch to communicate intention or attitude, but a lack of intuitive understanding for how this is done might lead speakers with ASD to either ignore this aspect of speech, or to use it in an atypical fashion. Of course, the degree to which speakers are more or less monotone or dynamic varies in neurotypical speakers as well.

As mentioned above, the speaker most consistently rated by both expert and naïve raters as atypical was one of the three speakers with ASD in Community 2. This speaker's speech is indeed quite dynamic in terms of pitch variation, and the peaks and valleys in pitch subjectively seem to fall in about the same places as in neurotypical speakers' productions of the same sentences (consistent with findings in Geelhand, Papastamou & Kissine, 2021). However, there

is an exaggerated and distinctive quality to this speaker's use of pitch variation. Furthermore, another salient aspect of this speaker's distinctive speech was *not* among the acoustic features measured in this study; several vowels were produced with an unusual quality, sounding like phonemes produced by a speaker of English as a second language. Indeed, broad measures such as pitch variation and articulation rate cannot capture the many small but important differences in sentence intonation that can convey critical information to the hearer.

How should this sort of subjective assessment of participants' speech be incorporated into our models? If vowel quality is of interest, then formant measures should be included into the acoustic profile. Alternatively, the atypicality of this speaker's speech could reflect an interaction of vowel quality with pitch variation and timing, whereas a different speaker's speech might be rated as similarly atypical, reflecting an entirely different combination of interacting features. The proposition of building a generalizable predictive model of the autistic voice reflects the enormous challenges in understanding the many-to-many mapping problem in the acoustics of speech (Liberman et al., 1967). We propose that all three possibilities described above (the inclusion of too few acoustic features, the misquantification of features, and the possibility that there is no single acoustically-definable profile of the autistic voice) present important challenges to building predictive models of atypical speech. They are not mutually exclusive, and all require further investigation. How researchers choose to address them will be in part based on the goals guiding the research.

An important aspect of ensuring progress on these complicated questions is the open sharing of data and code. While sharing raw audio files presents important data privacy problems, sharing anonymized acoustic features, as well as the code used for extracting and analyzing those features will be crucial for advancing research in this domain. To this end, we

provide full data and code at the online repository OSF: https://osf.io/zw67n/?view_only=63441e15ec264184bb7b0b22c4140a22.

Limitations and future research

The present study is exploratory in nature, and as such, there are several obvious avenues for future research. Among the most important next steps will be expanding the number and variety of features used to model the speakers' productions. As mentioned in the introduction, sentence-level pitch variation alone is a poor measure of intonation. Emphasizing a word in conversation, for example, can be achieved with a change of pitch, but also with a change in intensity, the addition of a pause, the lengthening or shortening of a vowel, or a combination of all of these. In addition, while scripted productions like the ones used in our study are useful because they are easily comparable, they are clearly less than ideal as a means of eliciting conversational prosody. Another obvious avenue for future research is thus the extension of the methods suggested here to more naturalistic data, including conversational data. Not only is prosody more vital for conducting successful conversations than it is for producing utterances in the lab, but the artificial setting of our elicitation may have impacted the speakers in a variety of ways, including the stress of having to "perform". A third opportunity for improvement will be obtaining higher-quality recordings, reducing background noise, and being more discerning in setting individualized windows for pitch extraction. Finally, it will be important to expand this research to larger, more diverse samples (including a wider range of language abilities, and speakers more diverse in sociodemographic factors) and to include other languages. Unfortunately, we expect that it will be difficult to achieve all of these within the context of a single study. For example, more naturalistic data will likely result in noisier recordings, larger samples will make careful, individualized feature extraction more difficult, and different

languages may make different use of prosodic features, making generalization difficult. Despite these difficulties, we find that the present results suggest the need to think in a more nuanced way about when, how, and whether the speech of people with ASD diverges from that of TD speakers, especially in conversation.

When de Marchena et al (2017) asked clinicians to identify the behavioral features they most associated with 'frank' ASD, atypical prosody was among the most common features mentioned, and they propose the pursuit of a 'narrow frank feature (e.g. avoidance of eye contact, or unusual prosody)' as an effective way to study and understand frank presentation autism (pg. 660). While our data suggest a viable means to explore and quantify these prosodic differences, they also point to the fact that even the relatively 'narrow' feature of prosody may contain multitudes.

Conclusion

Our main goal in this investigation was to test the use of a network-based community-detection algorithm in investigating atypical prosody and voice quality in ASD. Using four acoustic features, we built a network of speakers which reliably consisted of three communities: speakers that were more acoustically alike within their community and more dislike speakers outside their community. While two of the communities tended toward either more ASD or more NT speakers, the third community was populated by a nearly even split of NT and ASD speakers. Both expert and naïve raters were highly successful in identifying autistic speakers, regardless of which acoustic community the speakers had been placed in by the algorithm. Although we regard this exploratory study as a first proof of concept, we suggest that using network analysis to identify clusters of acoustically-similar speakers may lead to important insights into how atypical prosody, while always atypical, may be different in different ways.

Different in different ways

References

Al-Qatab, B. A., & Mustafa, M. B. (2021). Classification of dysarthric speech according to the severity of impairment: An analysis of acoustic features. IEEE Access, 9, 18183-18194.

Asperger, H. (1991). "*Autistic psychopathy" in childhood*. In U. Frith (Ed.), Autism and Asperger syndrome (pp. 37–91). Cambridge: Cambridge University Press. (Original work published 1944)

American Psychological Association (2013). *Diagnostic and statistical manual of mental disorders (DSM-5)*. American Psychiatric Pub.

Berends, M. S., Luz, C. F., Friedrich, A. W., Sinha, B. N., Albers, C. J., & Glasner, C. (2021). AMR-An R Package for working with antimicrobial resistance data. BioRxiv, 810622.

Boersma, P., & Weenink, D. (2001). *Praat, a system for doing phonetics by computer*.

Bone, D., Lee, C.-C., Black, M. P., Williams, M. E., Lee, S., Levitt, P., & Narayanan, S. (2014). The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1162–1177.

Bottalico, P., Codino, J., Cantor-Cutiva, L. C., Marks, K., Nudelman, C. J., Skeffington, J., Shrivastav, R., Jackson-Menaldi, M. C., Hunter, E. J., & Rubin, A. D. (2020). Reproducibility of Voice Parameters: The Effect of Room Acoustics and Microphones. Journal of Voice, 34(3), 320–334. https://doi.org/10.1016/j.jvoice.2018.10.016

Borrie, S. A., Barrett, T. S., Liss, J. M., & Berisha, V. (2020). Sync pending: Characterizing conversational entrainment in dysarthria using a multidimensional, clinically informed approach. Journal of Speech, Language, and Hearing Research, 63(1), 83-94.

Brockmann-Bauser, M., Beyer, D., & Bohlender, J. E. (2014). Clinical relevance of speaking voice intensity effects on acoustic jitter and shimmer in children between 5;0 and 9;11 years. International Journal of Pediatric Otorhinolaryngology, 78(12), 2121–2126. https://doi.org/10.1016/j.ijporl.2014.09.020

Bryant, G. A., Fessler, D. M., Fusaroli, R., Clint, E., Aarøe, L., Apicella, C. L., … others. (2016). Detecting affiliation in co-laughter across 24 societies. *Proceedings of the National Academy of Sciences*, *113*(17), 4682–4687.

Cappadocia, M. C., Weiss, J. A., & Pepler, D. (2012). Bullying experiences among children and youth with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *42*(2), 266–277.

Chen, L., Ravichandran, V., & Stolcke, A. (2021). Graph-based label propagation for semi-supervised speaker identification. arXiv preprint arXiv:2106.08207.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E, 70*(6), 066111.

Constantino, J. N., Davis, S. A., Todd, R. D., Schindler, M. K., Gross, M. M., Brophy, S. L., … Reich, W. (2003). Validation of a brief quantitative measure of autistic traits: Comparison of the social responsiveness scale with the autism diagnostic interview-revised. *Journal of Autism and Developmental Disorders*, *33*(4), 427–433.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*(5), 1–9.

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.

Different in different ways

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. Language and speech, 40(2), 141-201.

Dahlgren, S., Sandberg, A. D., Strömbergsson, S., Wenhov, L., Råstam, M., & Nettelbladt, U. (2018). Prosodic traits in speech produced by children with autism spectrum disorders–Perceptual and acoustic measurements. *Autism and Developmental Language Impairments, 3*, 2396941518764527.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). *Covarep - a collaborative voice analysis repository for speech technologies* (pp. 960–964). IEEE.

Diehl, D. B., Joshua J. Watson. (2009). An acoustic analysis of prosody in high-functioning autism. *Applied Psycholinguistics*, *30*(3), 385–404.

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

Djelantik, A. M. J., Robinaugh, D. J., Kleber, R. J., Smid, G. E., & Boelen, P. A. (2020). Symptomatology following loss and trauma: Latent class and network analyses of prolonged grief disorder, posttraumatic stress disorder, and depression in a treatment-seeking trauma-exposed sample. *Depression and Anxiety*, *37*(1), 26–34.

Drugman, T., & Alwan, A. (2019). Joint robust voicing detection and pitch estimation based on residual harmonics. arXiv preprint *arXiv*:2001.00459.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. Retrieved from http://www.jstatsoft.org/v48/i04/

Eskenazi, L., Childers, D. G., & Hicks, D. M. (1990). Acoustic correlates of vocal quality. *Journal of Speech, Language, and Hearing Research*, *33*(2), 298–306.

Filipe, M. G., Frota, S., Castro, S. L., & Vicente, S. G. (2014). Atypical prosody in Asperger syndrome: Perceptual and acoustic measurements. *Journal of Autism and Developmental Disorders, 44*(8), 1972-1981.

Fusaroli, R., Bang, D., & Weed, E. (2013). Non-linear analyses of speech and prosody in Asperger's syndrome. *International meeting for autism research*.

Fusaroli, R., Grossman, R., Bilenberg, N., Cantio, C., Jepsen, J. R. M., & Weed, E. (2021). Toward a cumulative science of vocal markers of autism: A cross-linguistic meta-analysis-based investigation of acoustic markers in American and Danish autistic children. *Autism Research*, 1-12. doi: 10.1002/aur.2661

Fusaroli, R., Grossman, R., Cantio, C., Bilenberg, N., & Weed, E. (2015). *The temporal structure of the autistic voice: A cross-linguistic investigation*. Poster session presented at the *International Meeting for Autism Research*.

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 Research*, *10*(3), 384–407.

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Geelhand, P., Papastamou, F., & Kissine, M. (2021). How do autistic adults use syntactic and prosodic cues to manage spoken discourse?. *Clinical Linguistics and Phonetics, 35*(12), 1184-1209.

Grossman, R. (2015). Judgments of social awkwardness from brief exposure to children with and without high-functioning autism. *Autism*, *19*(5), 580–587.

Hillenbrand, J., & Houde, R. A. (1994). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, *39*(2), 311–321.

Huang, D. Z., Minifie, F. D., Kasuya, H., & Lin, S. X. (1995). Measures of vocal function during changes in vocal effort level. Journal of Voice, 9(4), 429–438. https://doi.org/10.1016/S0892-1997(05)80206-1

Kissine, M., Geelhand, P., Philippart De Foy, M., Harmegnies, B., & Deliens, G. (2021). Phonetic inflexibility in autistic adults. *Autism Research, 14*(6), 1186-1196.

Kissine, M. & Geelhand, P. (2019). Brief Report: Acoustic Evidence for Increased Articulatory Stability in the Speech of Adults with Autism Spectrum Disorder. J. Autism Dev. Disord. 49, 2572–2580

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Kreiman, J., & Gerratt, B. (2010). Perceptual sensitivity to first harmonic amplitude in the voice source. *The Journal of the Acoustical Society of America*, *128*(4), 2085–2089.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431.

Different in different ways

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., … Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223.

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.

Marchena, A. de, & Miller, J. (2017). "Frank" presentations as a novel research construct and element of diagnostic decision-making in autism spectrum disorder. *Autism Research*, *10*(4), 653–662.

Maslov, S., & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science, 296*(5569), 910-913.

Mayo, J. (2015). Prosodic Phrasing in Adolescents with High Functioning Autism: Production Following Intervention and Under Dual Load Conditions. Ph.D. Thesis, University of Connecticut.

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

Mittal, V., & Sharma, R. K. (2021). Machine learning approach for classification of Parkinson disease using acoustic features. Journal of Reliable Intelligent Environments, 7(3), 233-239.

Mesibov, G. B. (1992). Treatment issues with high-functioning adolescents and adults with autism. In *High-functioning individuals with autism* (pp. 143–155). Springer.

Different in different ways

Mottron, L., & Bzdok, D. (2020). Autism spectrum heterogeneity: Fact or artifact? *Molecular Psychiatry, 25*(12), 3178-3185.

Murphy, P. J. (2000). Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals. *The Journal of the Acoustical Society of America, 107*(2), 978-988.

Nadig, A., & Shaw, H. (2012). Acoustic and perceptual measurement of expressive prosody in high-functioning autism: Increased pitch range and what it means to listeners. *Journal of Autism and Developmental Disorders*, *42*(4), 499–511.

Nakai, Y., Takashima, R., Takiguchi, T., & Takada, S. (2014). Speech intonation in children with autism spectrum disorder. *Brain and Development*, *36*(6), 516–522.

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.

Parola, A., Simonsen, A., Lin, J. M., Zhou, Y., Huiling, W., Ubukata, S., ... & Fusaroli, R. (2022). Voice patterns as markers of schizophrenia: building a cumulative generalizable approach via cross-linguistic and meta-analysis based investigation. medRxiv.

Peng, C., Chen, W., Zhu, X., Wan, B., & Wei, D. (2007). Pathological voice classification based on a single Vowel's acoustic features. In 7th IEEE International Conference on Computer and Information Technology (CIT 2007) (pp. 1106-1110). IEEE.

Peppé, S. J. E. (2009). Why is prosody in speech-language pathology so difficult. *International Journal of Speech and Language Pathology*, *11*(4), 258–271.

Peppe, S., McCann, J., Gibbon, F., O'Hare, A., & Rutherford, M. (2007). Receptive and expressive prosodic ability in children with high-functioning autism. *Journal of Speech, Language and Hearing Research*, *50*(4), 1015-1028.

Quene H (2022). _hqmisc: Miscellaneous Convenience Functions and Dataset. R package version 0.2-1, https://CRAN.R-project.org/package=hqmisc

R Core Team. (2022). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Redford, M. A., Kapatsinski, V., & Cornell-Fabiano, J. (2018). Lay listener classification and evaluation of typical and atypical children's speech. *Language and Speech*, *61*(2), 277–302.

Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. *arXivPhys. Rev. E 74 (2006) 016110*, 0603718v1.

Roid, G. H. (2003) *Stanford-Binet Intelligence Scales: Fifth Edition*. Itasca, IL: Riverside

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).

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. Journal of psycholinguistic research, 25(2), 193-247.

Shriberg, L. D., Paul, R., Black, L. M., & Santen, J. P. van. (2011). The hypothesis of apraxia of speech in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *41*(4), 405–426.

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*.

Different in different ways

Shriberg, L. D., & Widder, C. J. (1990). Speech and prosody characteristics of adults with mental retardation. *Journal of Speech, Language, and Hearing Research*, *33*(4), 627–653.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., ... & Evanini, K. (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In 17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5 (pp. 2001-2005).

Team, R. C. (2018). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Traag, V. A., & Bruggeman, J. (2008). Community detection in networks with positive and negative links. *arXivPhys. Rev. E 80, 036115, (2009)*, 0811.2329v3.

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.

Wang, J., Xiao, X., Wu, J., Ramamurthy, R., Rudzicz, F., & Brudno, M. (2021). Speaker attribution with voice profiles by graph-based semi-supervised learning. arXiv preprint arXiv:2102.03634.

Waterhouse, L. (2013). *Rethinking autism: Variation and complexity*. Academic Press.

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*, *63*(6), 1762–1775.

Wiig, E. H., Semel, E. M., & Secord, W. (2003). *CELF 5: Clinical Evaluation of Language Fundamentals*. Pearson/PsychCorp.

Wolfe, V., Fitch, J., & Martin, D. (1997). Acoustic measures of dysphonic severity across and within voice types. *Folia Phoniatrica Et Logopaedica*, *49*(6), 292–299.

Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, *71*(6), 1544–1550.

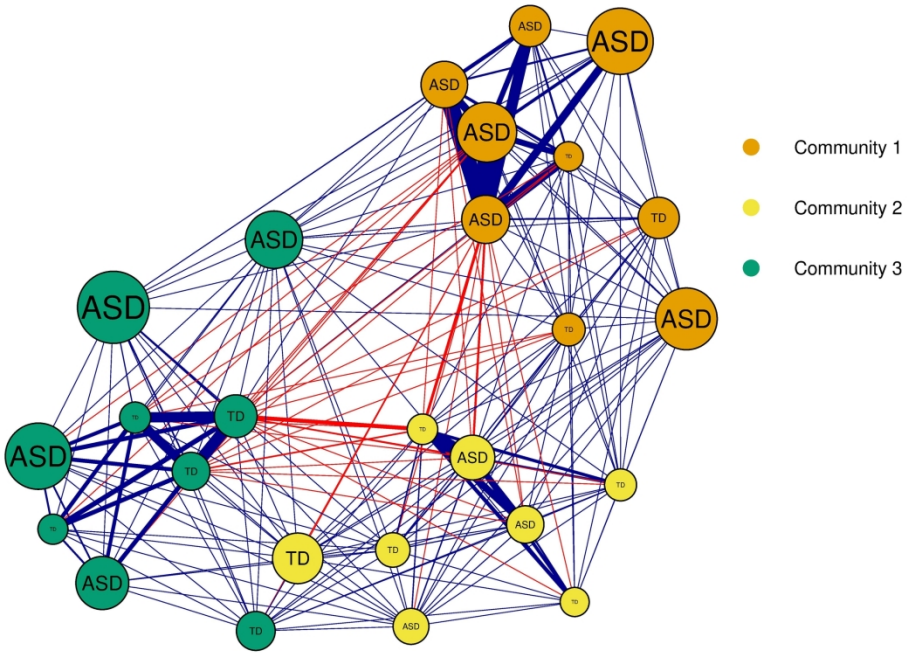


Figure 1: Network of acoustically-defined communities. Each node represents a single speaker. Blue connections between nodes represent positive correlations, red lines represent negative correlations. Line thickness indicates the absolute strength of the correlation.

210x181mm (300 x 300 DPI)

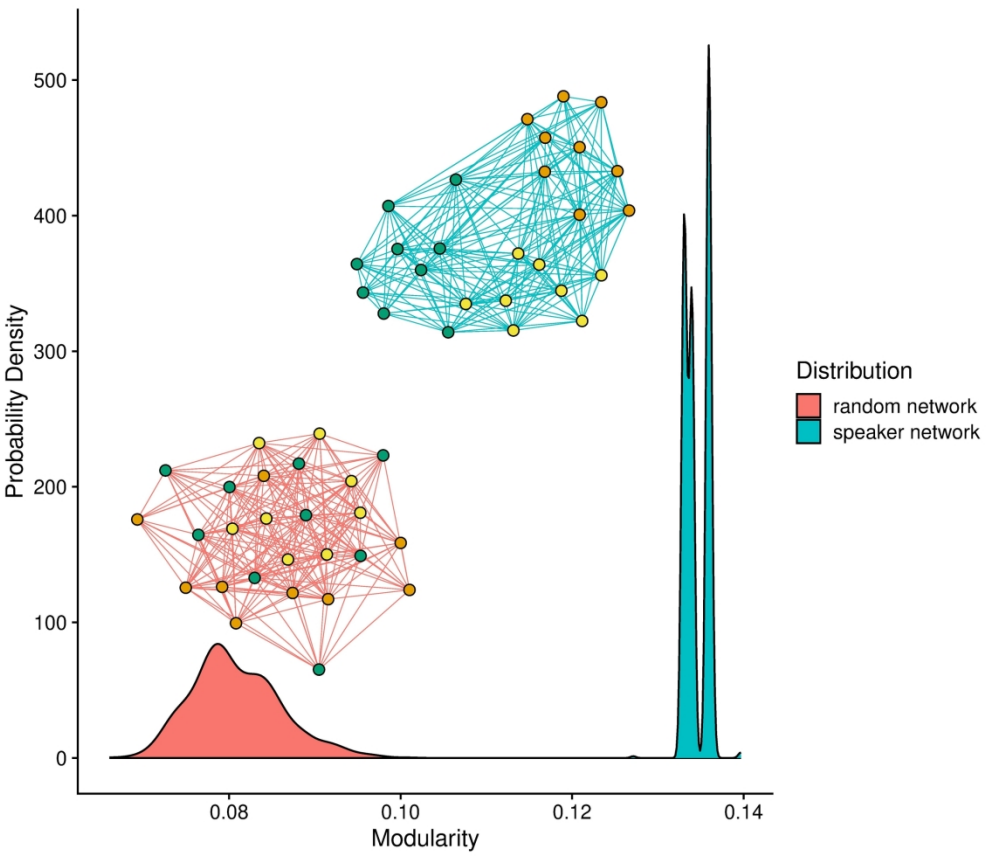


Figure 2: Distributions of modularity scores for 1000 random networks built with the same constraints as the actual network of speakers, and the distribution of modularity scores for 1000 runs of community detection on the speaker network. Also displayed are a sample random network and the actual speaker network. The colors of the connections in the sample networks indicate which distribution they are drawn from. The colors of the nodes indicate community membership.

210x181mm (300 x 300 DPI)

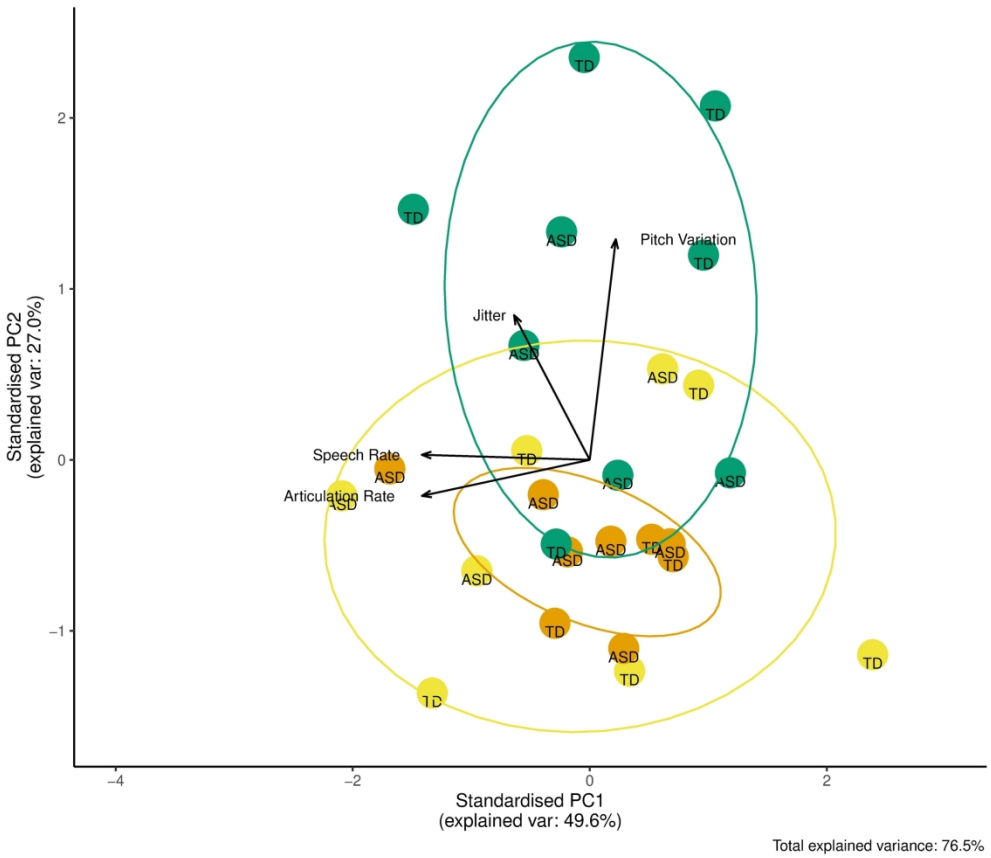


Figure 3: Speakers and communities visualized along the first two dimensions of a principal components decomposition. Ellipsis colors and point colors indicate communities detected by the spin glass algorithm (orange = Community 1, yellow = Community 2, green = Community 3). Arrows indicate the influence of each of the features on the components.

210x181mm (300 x 300 DPI)

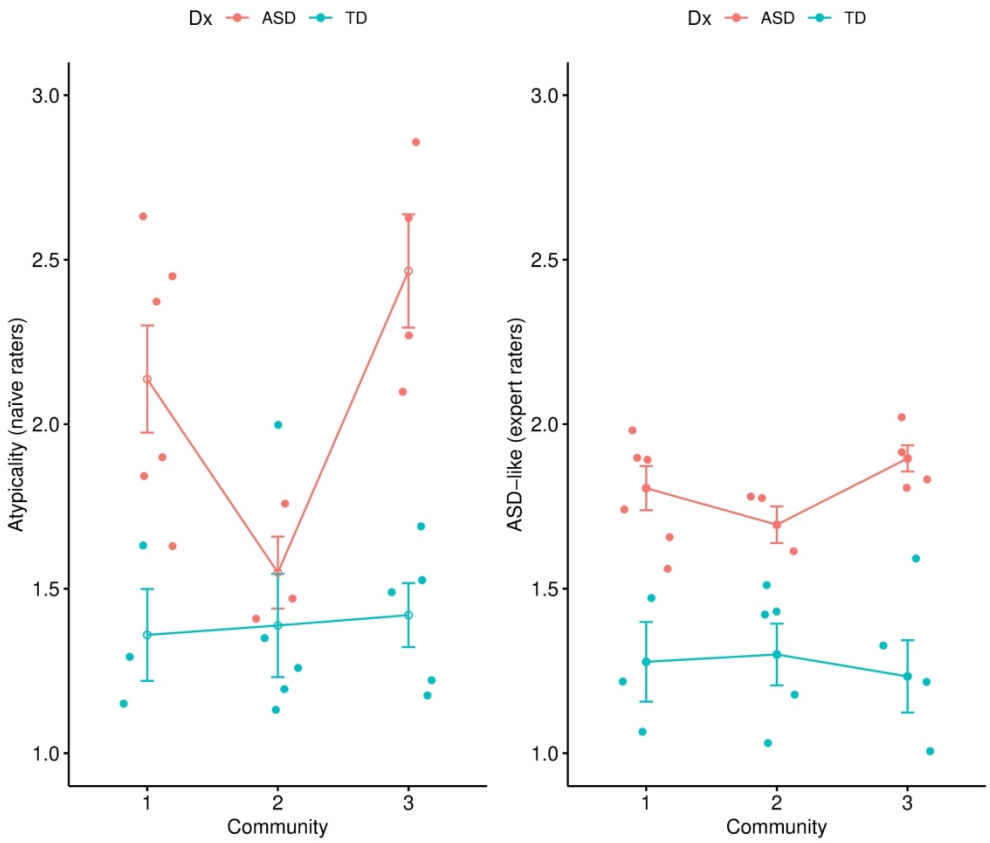


Figure 4: Relationship between ratings by naïve (left) and clinically-trained (right) raters and community membership for ASD (red) and NT (blue) participants

210x181mm (300 x 300 DPI)