Acoustic-Prosodic Correlates of 'Awkward' Prosody in Story Retellings from Adolescents with Autism

Daniel Bone[1], *Matthew P. Black*[1,2], *Anil Ramakrishna*[1], *Ruth Grossman*[3,4], *Shrikanth Narayanan*[1,2]

[1]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA, USA
[2]Information Sciences Institute, USC, Marina del Rey, CA, USA
[3]Department of Communication Sciences and Disorders, Emerson College, Boston, MA, USA
[4]4University of Massachusetts Medical School Shriver Center, Worcester, MA, USA

dbone@usc.edu, http://sail.usc.edu

Abstract

Atypical speech prosody is a primary characteristic of autism spectrum disorders (ASD), yet it is often excluded from diagnostic instrument algorithms due to poor subjective reliability. Robust, objective prosodic cues can enhance our understanding of those aspects which are atypical in autism. In this work, we connect objective signal-derived descriptors of prosody to subjective perceptions of prosodic *awkwardness*. Subjectively, more *awkward* speech is less *expressive* (more monotone) and more often has perceived awkward rate/rhythm, volume, and intonation. We also find *expressivity* can be quantified through objective intonation variability features, and that speaking rate and rhythm cues are highly predictive of perceived *awkwardness*. Acoustic-prosodic features are also able to significantly differentiate subjects with ASD from typically developing (TD) subjects in a classification task, emphasizing the potential of automated methods for diagnostic efficiency and clarity.

Index Terms: prosody, autism spectrum disorders, rhythm, intonation, perceived awkwardness

1. Introduction

Speech prosody—the rhythm, stress, and intonation of speech—plays a critical role in effective communication, disambiguating meaning and conveying paralinguistic information like attitude and emotion. Prosodic patterns differ among speakers, and a listener must take into account a particular level of variability; but at a certain point, a threshold is crossed from acceptable variability to perceived abnormality.

Atypical speech prosody is a primary symptom of autism spectrum disorder (ASD), a development disorder defined by impairments in social communication and reciprocity, as well as restricted, repetitive behavioral patterns and interests [1]. Prosodic deficits in ASD, which exist in both perception and production [2], have a detrimental impact on an individual's social-communicative abilities. People with ASD generally have difficulty discerning a speaker's intent from prosody, and their own speech is often perceived as awkward. These prosodic abnormalities may be attributable to impairments in Theory of Mind (the ability to decode another person's mental state, [3]).

Speech prosody is regarded as a high-impact, understudied area in autism [4]; little is known about the specific prosodic abnormalities within ASD and their prevalences. While 'Speech Abnormalities Associated with Autism (Intonation/Volume/Rhythm/Rate)' is coded in the gold-standard Autism Diagnostic Observation Scale (ADOS, [5]), it is not included in the diagnostic algorithm due to subjective disagreements between clinicians. Objective computational methods can fill the gap [6]; atypical prosody quantification can support advances in the understanding and treatment of autism—from better stratification which aids neuro-biological research into ASD etiology, to better prosodic intervention systems.

Research on atypical prosody in autism has largely focused on human perception of read or spontaneous speech. Such studies have reported, for example, atypicalities in sentential [7] and contrastive stress [8], increased pausing [9], and abnormal voice quality [10]. A perceptual rating tool, the Prosody-Voice Screening Profile (PVSP, [11]), has been used to assess global prosodic attributes in spontaneous speech, finding more inappropriate phrasing, stress, and resonance for ASD subjects [12].

Scalable acoustic correlates of atypical prosody are relatively unexplored; such studies have reported increased f0 variability [13], prosodic differences in stress production [14], and higher maximum f0 for ASD subjects [15]. Our previous work related a variety of interpretable, automatically-extracted prosodic (intonation, volume, rate, rhythm) and language cues in a diagnostic session to ASD severity. Not only were the child's cues predictive of their level of symptom severity, but so were those of the psychologist, who must continually adjust her behavior to that of the child's throughout the interaction [16, 6, 17]. However, not all people with ASD have prosodic difficulties, so it is desirable to relate our acoustic measures directly to perceived *atypical* prosody.

In this study, we ask naive human raters to assess various types of prosodic awkwardness, then link these perceptions to objective acoustic-prosodic measures. Raters score overall *awkwardness*, as well as awkwardness of individual components of prosody: rate/rhythm, volume, and intonation/stress. We expect that agreement will be highest at the cumulative level, that something sounds "odd." Vocal *expressivity* is also rated since ASD prosody is described as monotone or overly exaggerated, which may contribute to perceived awkwardness. Through this work, we aim to enhance our understanding of signal-derived speech prosody measures, which are vital to behavioral interaction analyses [6] and the creation of automated clinical tools.

2. Methodology

In the following sections we discuss the data collection and participant demographics, perceptual rating scheme, acoustic-prosodic features, and machine learning data analysis.

2.1. Data Collection and Participants

Data were recorded as part of an affective story retelling task. Participants initially viewed a stimulus video in which an actor, "Safari Bob", stated that he needed someone to fill in for

Table 1: *Participant demographics. '$*$' designates difference at α=0.05 level by Wilcoxon rank-sum test.*

	N	*Age (yr.)*	*Female*	*V-IQ*	*P-IQ*	*Rec. Vocab.*	*Reading*
ASD	43	12.9	1 (2.3%)	105	102*	111*	105
TD	26	13.6	2 (7.7%)	112	113*	122*	108

him as host of a children's television show. Participants listened to a story told by Safari Bob, then retold it with the story text displayed on screen. We focus our preliminary analyses on one of the four stories, "Elephants", which contains five sentences. Of the possible 345 utterances for analysis (69x5), 322 are selected (μ=4.7s, σ=1.7s) post-exclusion of poor audio quality and utterances that went far off-script. ASD and TD (typically developing) participant demographics are presented in Table 1, including: verbal IQ, performance IQ, receptive vocabulary (as measured by the Peabody Picture Vocabulary Test-Revised [18]), and reading level (as measured by the Woodcock Johnson Test [19]). We control for the statistically-significant group differences in demographics during later analysis. More details can be found in the primary paper on this database [15].

2.2. Perceptual Ratings of Prosody

Our study is motivated by the general perceptions of awkwardness that occur when interacting with individuals with autism; naive raters of prosody are able to detect an overall quality of "awkwardness" in an ASD individual's speech [20].

Each utterance is scored on N-point Likert scales by 15 naive raters on Amazon Mechanical Turk (MTurk). Raters could listen to the files multiple times while judging a speaker's overall *awkwardness* and other related constructs. Specifically, 'Awkwardness' is obtained through inversion of a 'Naturalness' (non-awkwardness) rating, which is on a 4-point scale from 'Very Awkward' to 'Natural (Not Awkward).' Also, raters mark the presence of awkwardness (binary) in three components of prosody—'Rate/Rhythm', 'Volume', and 'Intonation/Stress'. Lastly, 'Expressivity' (animation) is rated on a 5-point scale from 'Extremely Flat or Monotone' to 'Overly Animated.'

Final ratings are obtained through averaging scores per utterance. Given the variable quality of raters from MTurk, we remove raters with very poor agreement with the initial mean (ρ_S<0.2) and raters who evaluated less than 10 utterances. Additionally, awkward prosody component scores (binary) are z-normalized per-rater before fusion, which improves agreement.

2.3. Acoustic-Prosodic Features

Prosodic atypicalities associated with autism have been reported in the domains of intonation, volume, rate, and voice quality; as such, we compute a total of 37 features which target these qualitative constructs. A novelty of this work is that we compare features that jointly model prosody as it occurs with exact lexical content versus those that do not. Our utterance-level features that do not precisely model the lexical content are grouped as *rate & rhythm*, *voice quality*, and *intonation*. Features which model prosody jointly with lexical content include *exemplar-based intonation/stress* and *transcript-matching* cues.

Speech *rate & rhythm* comprise 12 cues. Speech articulation *rate* is measured as the median and inter-quartile ratio (IQR) of individual syllable rates per-utterance; syllable boundaries are determined by forced-alignment using HTK. Speech *rhythm*, the temporal patterning of speech units, is quantified using Pairwise Variability Indices (PVI, [21]) and Global Interval Proportions (GIP, [22, 23]). Pairwise variability indices measure durational variability of adjacent linguistic units; we compute normalized and unnormalized PVI measures for consonants, vowels, and syllables. GIP features include the percentage of vowel speech and the standard deviations of vowel and consonant durations. We also compute the percentage of pausing within an utterance, a key facet of rhythm.

Voice quality is captured by six features: median and IQR of syllabic jitter, shimmer, and harmonics-to-noise-ratio (HNR). Jitter and shimmer, the local variability in pitch and intensity, respectively, are calculated using the method described in [6], utilizing Praat [24]. HNR is extracted using VoiceSauce [25].

We model *intonation* through syllable-level parametrization of pitch and intensity signals. We compute the slope and curvature of these signals per-syllable, then calculate utterance-level median and IQR. Raw signal means and standard deviations are also extracted, totaling 12 features. This technique may capture speaker idiosyncrasies in intonation.

Using *exemplar-based template* features, we implicitly model *intonation & stress* as they occur jointly with the lexical content. These features have previously been used in children's read prosody assessment computed against an adult narration exemplar [26], and were previously proposed by Bone et al. for studying prosody in ASD [27]. *Exemplar* features model the evolution of a prosodic contour with spoken words compared to a reference; an example for duration is shown in Fig. 1. First, prosodic contours are extracted (pitch, intensity, and duration), which are then time-aligned with word-boundaries. For each word and contour, a single feature functional is computed (we use median), producing a representation in which each word holds a single prosodic value. Next, we obtain a single exemplar for comparison by averaging five productions with the best rating (e.g., least awkward). This is done per-rating in a leave-one-subject-out fashion; in the case of predicting ASD diagnosis, we use all TD subjects for deriving the exemplar. Lastly, we compute the Spearman's correlation (ρ_S) between an observed prosodic template and the exemplar, generating one feature per prosodic signal. Missing words or feature values are penalized; we scale ρ_S by the percentage of valid feature values.

Transcript-matching features relating observed and reference transcripts include percentages of correct, inserted, deleted (Fig. 1), and substituted words (computed via NIST SCTK).

2.4. Statistical Analysis and Machine Learning

Two types of analyses are conducted: correlation and prediction/classification. Support Vector Regression and logistic regression [28] are performed in a speaker-independent/sentence-

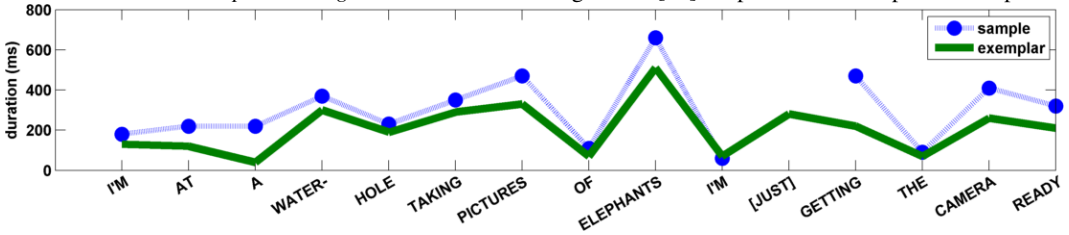

Figure 1: *Duration sample versus exemplar for one sentence, where "just" was missing from the sample production. Computation: ρ_S=0.90; fraction of samples with a feature value=$\frac{14}{15}$; final* ***exemplar-score***$=0.90 * \frac{14}{15}=$***0.84*** *.*

Table 2: *Spearman's ρ inter-rater reliability (***sig.** *at α=0.05).*

Code	*Expr*	*Awk*	*R/R*	*Vol*	*Inton*
Spearman's ρ	**0.70**	**0.57**	**0.42**	**0.37**	**0.25**

independent manner to support generalization of results; parameters are tuned using two-level nested cross-validation. Perceptual ratings may vary for utterances from an individual speaker, but their ASD diagnosis is constant. Therefore, analysis between subjective ratings and acoustics is conducted for all utterances (treated individually); in contrast, we pool (average) samples when predicting autism diagnosis; pooling predictions across utterances also models realistic application of an automatic system. This topic is discussed further in Section 4.2.

3. Analysis of Perceptual Ratings

In this section we discuss the inter-rater reliability of different perceptual codes by our naive raters (Table 2), as well as correlations between perceptual ratings, demographic variables (receptive vocabulary, P-IQ, & age), and ASD diagnosis (Table 3). *Table Legend*: *Expr* - expressivity; *Awk* - awkwardness; awkward *R/R* - rate/rhythm, *Vol* - volume, *Inton* - intonation/stress.

The naive raters achieve moderate or substantial agreement for overall *awkardness* and *expressivity* (calculated as the median Spearman's correlation between each rater and the mean score of the other evaluators that rated those utterances). Understandably, there is much lower agreement for the more specific components of prosodic awkwardness, listed in descending order as follows: *rate/rhythm*, *volume*, and *intonation/stress*. Cumulative perceptions tend to have much higher agreement than more specific items; this is true for autism diagnostic instruments [5, 29] and for the PVSP prosody examination [11].

Correlations between different perceptual ratings can also inform the human perceptual process. Speakers that are perceived as more generally *awkward* are also heard as less *expressive* (ρ_S=−0.39), or more *monotone*. The global perception of awkwardness can be further decomposed; awkward speakers tend to have more awkward *rate/rhythm* (ρ_S=0.80), awkward *intonation* (ρ_S=0.50), and awkward *volume* (ρ_S=0.35); appropriate timing, or *rate/rhythm*, is a critical factor in judging overall *awkwardness* for an utterance.

Next, we consider dependencies between perceptual codes and demographic variables. Subjects with higher receptive vocabulary are perceived as more *expressive* (ρ_S=0.27) and generally less *awkward* (ρ_S=−0.39)—this specifically includes the domains of awkward *rate/rhythm* (ρ_S=−0.37) and awkward *intonation* (ρ_S=−0.44); very similar relations exist between performance IQ and perceptual ratings. Younger subjects tend to have higher incidence of awkward *volume* (ρ_S=−0.29).

A primary goal of this study is to examine prosodic awkwardness in autism. ASD subjects' speech was perceived as different than control subjects' speech, even though the MTurk raters were blind to study purpose, demographic makeup, and task description. ASD speech was perceived as more *awkward* (ρ_S=0.50), and had a higher rate of awkward *rate/rhythm* (ρ_S=0.48), *volume* (ρ_S=0.32), and *intonation* (ρ_S=0.33). All relations with ASD diagnosis remain significant (α=0.05) after controlling for demographics (receptive vocab., P-IQ, & age).

Table 3: *Correlations between speaker-averaged ratings, demographics, and ASD diagnosis.* **Bolded** *implies sig. at α=0.05.*

	Awk	*R/R*	*Vol*	*Inton*	*Vocab*	*P-IQ*	*Age*	*ASD*
Expr	**-0.39**	**-0.25**	0.09	0.16	**0.27**	**0.31**	0.05	-0.06
Awk		**0.80**	**0.35**	**0.50**	**-0.39**	**-0.36**	0.12	**0.50**
R/R			0.07	**0.41**	**-0.37**	**-0.30**	-0.14	**0.48**
Vol				-0.05	0.07	-0.10	**-0.29**	**0.32**
Inton					**-0.44**	-0.08	0.06	**0.33**
Vocab						**0.40**	-0.05	-0.22
P-IQ							-0.08	**-0.27**
Age								-0.10

4. Acoustic-Prosodic Cues of Awkwardness

Interpretable, objective signal measures can provide a bottom-up explanation for these human perceptions of prosody, allowing scalable application to larger data. In section 4.1, the most informative prosodic cues for each perceptual rating are discussed, and in section 4.2, prosodic cues are used to predict perceptual ratings and autism diagnosis.

4.1. Correlational Feature Analysis

The top five features related to each perceptual rating and ASD diagnosis are provided in Table 4. Since the value of a feature is dependent on various sources of noise in the feature extraction process, we cannot confidently state that a construct is uninformative of a target variable, only that the extracted feature is not.

The top cues for overall *awkwardness* relate primarily to timing; speech that is less awkward has less pausing, less local variability in vowel duration, and a higher correlation with the syllable-duration exemplars. Less awkward-sounding productions also adhere more to the transcript and insert fewer words. Since raters were not given the text, aberrations from the transcript likely should not factor into the ratings, but may have had other prosodic effects (e.g., increased pausing) which are more directly relevant to perceived awkwardness.

Although agreement on awkwardness in prosodic subcomponents is lower, acoustic-prosodic cues can still provide insights into the human perceptual process. Perceived awkward *rate/rhythm* is captured by divergence from the normal relative word-duration (exemplar features), increased pausing, and more variable syllabic-duration. Awkward use of *volume* correlates with increased pausing, higher median HNR, and less variable syllabic-intensity slope—a possible marker of the flat, stilted expression seen in ASD. Awkward *intonation/stress* is best explained by slower articulation rate (syl/s), as well as more variable duration and syllabic-intensity dynamics.

Perceived *expressivity* is best captured by the variability of pitch and intensity contours. More expressive speech has more variable pitch and intensity, higher and more variable jitter, and less pausing—all indicators of higher vocal arousal [30]. These findings support the use of dynamic-intonation variability measures to assess monotone intonation in ASD.

Acoustic-prosodic cues also serve as evidence of differences between speech from ASD and TD subjects. All presented features are at least marginally significant (α=0.10

Table 4: *Top five features correlated (ρ_S) with perceptual ratings and ASD diagnosis (ASD≡1, TD≡0).* **Bold***: p<0.01; else p<0.05.*

	Perceptual Rating					**Diagnosis**
Ranking	*Expressivity*	*Overall Awkwardness*	*Awkward Rate/Rhy.*	*Awkward Volume*	*Awkward Intonation*	*Autism Spectrum*
Feat. 1	**0.43** f0 σ	**0.47** pause %	**-0.40** dur. model ρ	0.28 pause %	**0.42** PVI vowels	**-0.40** dur. model ρ
Feat. 2	**0.42** jitter Mdn	**-0.39** dur. model ρ	**0.39** pause %	-0.25 int. slope IQR	**0.37** vowel dur. σ	**0.32** pause %
Feat. 3	**0.40** jitter IQR	**0.36** PVI vowels	**0.35** vowel dur. σ	0.24 HNR Mdn	**0.35** int. slope IQR	-0.30 correct %
Feat. 4	-0.30 pause %	**-0.36** correct %	**0.32** PVI vowels	— —	**-0.34** rate Mdn (syl/s)	-0.27 rate IQR (syl/s)
Feat. 5	0.29 int. σ	**0.34** insertion %	0.31 PVI syllables	— —	**0.33** int. curv. IQR	0.26 substituted %

Table 5: *Regression and classification of perceptual ratings and ASD diagnosis via acoustic features and demographic variables.* **Bolded** *statistics are significant at the* α=0.05 *level by one-sided tests.* $N_{ratings}$=322, $N_{Diag.}$=69.

	Perceptual Rating					Diagnosis
Features	*Expr*	*Awk*	*RR*	*Vol*	*Into*	*ASD*
Baseline: Demog.	**0.16**	**0.32**	**0.25**	**0.22**	**0.20**	**63%**
Rate/Rhythm	**0.25**	**0.53**	**0.45**	**0.24**	**0.30**	**69%**
Exemplar	**0.15**	**0.41**	**0.40**	**0.13**	0.02	56%
Voice Qual.	**0.43**	**0.17**	0.08	**0.12**	-0.06	46%
Trans. Match	0.00	**0.23**	**0.24**	-0.14	0.04	59%
Intonation	**0.38**	0.06	0.02	**0.12**	**0.25**	48%
Feature Fusion	**0.55**	**0.56**	**0.47**	**0.21**	**0.36**	**65%**
Agreement	**0.70**	**0.57**	**0.42**	**0.37**	**0.25**	*N/A*
metric	Spearman's correlation (ρ_S)					UAR

level) after controlling for receptive vocab., P-IQ, and age, unless otherwise specified. ASD productions tend to have lower correlations with durational exemplars trained (speaker-independently) on TD subjects' speech—objective evidence of differences in ASD subjects' use of duration relative to lexical content. Interestingly, ASD subjects had less variable articulation rate (syl/s), another potential correlate of 'monotone' production. Speech from ASD subjects also contained more pauses, and also matched the transcript less often—i.e., more correct words and less substitutions, although substitution % becomes non-significant after accounting for demographics.

Many of the most informative signal cues pertain to timing, rate, and rhythm, which is unsurprising since the related subjective code is a highly-explanatory factor for overall perceived *awkwardness*; for example, pause frequency is an invaluable cue in assessing *awkwardness* (as in other automatic speech assessment scenarios such as children's literacy [31]). Although the agreement on perceived awkward *volume* and *intonation* is relatively low, several intuitive signal relations emerged.

4.2. Predicting Perceptual Ratings

While the individual correlational analysis in the previous section can inform interpretation in human-human behavioral interaction analyses, automatic systems that support clinical researchers can rely on joint modeling of many features. In this section, we analyze the performance of different feature categories in predicting ratings of prosody and autism diagnosis (Table 5); results of such experiments can inform not only the acoustical dependence of our perceptions, but also those cues which are associated with speech abnormalities of autism.

Rate & Rhythm features are significantly predictive of all perceptual ratings, producing the highest performances in nearly all experiments, with the sole exception being *expressivity*. The analysis of Section 4.1 showed timing features were excellent correlates of perceived *awkwardness*. In fact, Rate & Rhythm features alone meet or exceed the baseline results, and achieve performance on par with inter-rater agreement for all four awkward-prosody codes—e.g., for overall *awkwardness* ρ_S=0.53 versus an agreement of ρ_S=0.57.

Exemplar features, which measure dynamic-differences in word-time prosodic feature streams compared to a baseline model (exemplar), produce significant prediction for overall *awkwardness*, awkward *rate/rhythm*, and awkward *volume*, albeit below that of Rate & Rhythm features. The other lexical-modeling features, Transcript Matching statistics, are able to account for a small amount of the variance associated with overall *awkwardness* and awkward *rate/rhythm* through prediction.

The remaining two feature sets excel at quantifying *expressivity*. The global and local variance in f0 and intensity captured by Intonation and Voice Quality features predict *expressivity* ratings moderately well (ρ_S=0.38 and ρ_S=0.43, respectively). When Rate & Rhythm and the other acoustic cues are included in the model, prediction improves to ρ_S=0.55, still below inter-rater agreement (ρ_S=0.70). *Expressivity* is the only code for which there is a large gain from fusion over Rate & Rhythm features alone, highlighting the importance of timing cues.

Lastly, we predict *ASD* diagnosis using the provided acoustic-prosodic cues. Since autism diagnosis is an intricate procedure lasting hours and incorporating various sources of information, we should not expect to achieve very high performance from speech alone, much less from a few read utterances. Still, prediction from acoustics allows us to observe the importance of a group of signal cues in discovering differential patterns between groups (ASD and TD). Such findings can eventually lead to improved automatic assessment and monitoring systems or automatic prosodic tutor systems. Unweighted average recall (UAR) prediction performances are presented for ASD diagnosis (50% UAR is chance); speaker- and sentence-independent models are evaluated on each utterance, then predictions are aggregated through majority voting per speaker.

Rate & Rhythm features achieve the best performance in classifying *ASD* (69%), likely due to their utility in quantifying *awkward* prosody, which we showed in Section 3 to be associated with ASD diagnosis. Moreover, this is the only individual feature group which produces significant classification UAR[1]. After fusing all acoustic-prosodic features, classifier performance drops (potentially due to insufficient data size) to 65%. Demographic features (receptive vocab., P-IQ, and age), also achieve significant prediction at 63%.

5. Conclusion and Future Work

Speech cues are critical to finer characterization of autism spectrum disorder, yet there has been little headway toward a generalizable operational definition of prosodic atypicalities in ASD; e.g., prevalence estimates for various prosodic abnormalities are still unknown. Fortunately, speech processing can provide scalable, objective measures to support scientific advances.

In this work, we link acoustic-prosodic cues to general perceptions of speech prosody. Naive raters reach moderate to substantial agreement on cumulative aspects of prosody, but have lower agreement about components of prosody, highlighting the difficulty of explicating a general assessment; objective cues offer insights into that process. Rate & rhythm features are predictive of various awkwardness codes, producing correlations approaching inter-rater agreement; these timing cues additionally differentiate ASD and TD groups. Exemplar features, which jointly model prosody and lexical content, are also significantly informative of *awkwardness*. Lastly, dynamic intonation features can objectively quantify perceived *expressivity*.

In the future, we will continue to investigate the relationship between perception of prosody and acoustic cues. Acoustic-prosodic cues can provide novel insights into dyadic interactions involving children with autism [6]. Eventually, systems incorporating automatically extracted signal cues may be created for enhanced diagnostics and behavioral monitoring as well as prosodic intervention.

6. Acknowledgments

Work supported by NSF, NIH, and DoD, as well as ARCS and the USC Alfred E. Mann Innovation in Engineering Fellowship.

[1]Measured by a conservative one-sided binomial proportions test with N=2$*N_{minority-class}$ as described in [6].

7. References

[1] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders, (DSM-5®)*. American Psychiatric Pub, 2013.

[2] R. Paul, A. Augustyn, A. Klin, and F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 35, no. 2, pp. 205–220, 2005.

[3] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a theory of mind??" *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

[4] J. McCann and S. Peppe, "Prosody in Autism Spectrum Disorders: A Critical Review," *Int. J. Lang. Comm. Dis.*, vol. 38, pp. 325–350, 2003.

[5] C. Lord, S. Risi, L. Lambrecht, E. Cook, B. Leventhal, P. DiLavore, A. Pickles, and M. Rutter, "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*, vol. 30, pp. 205–223, 2000.

[6] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "The Psychologist as an Interlocutor in Autism Spectrum Disorder Assessment: Insights from a Study of Spontaneous Prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, pp. 1162–1177, 2014.

[7] R. Paul, L. D. Shriberg, J. McSweeny, D. Cicchetti, A. Klin, and F. Volkmar, "Brief Report: Relations between Prosodic Performance and Communication and Socialization Ratings in High Functioning Speakers with Autism Spectrum Disorders," *Journal of Autism and Developmental Disorders*, vol. 35, pp. 861–869, 2005.

[8] S. Peppe, J. McCann, F. Gibbon, A. O'Hare, and M. Rutherford, "Receptive and Expressive Prosodic Ability in Children with High-Functioning Autism," *Journal of Speech, Language, & Hearing Research*, vol. 50, pp. 1015–1028, 2007.

[9] R. B. Grossman, R. H. Bemis, D. P. Skwerer, and H. Tager-Flusberg, "Lexical and affective prosody in children with high-functioning autism," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 3, pp. 778–793, 2010.

[10] W. Pronovost, M. P. Wakstein, and D. J. Wakstein, "A longitudinal study of the speech behavior and language comprehension of fourteen children diagnosed atypical or autistic." *Exceptional children*, 1966.

[11] L. D. Shriberg, J. Kwiatkowski, C. Rasmussen, G. L. Lof, and J. F. Miller, "The prosody-voice screening profile (pvsp): Psychometric data and reference information for children," *The Prosody-Voice Screening Profile (PVSP): Psychometric data and reference information for children*, 1992.

[12] L. D. Shriberg, R. Paul, J. L. McSweeny, A. Klin, D. J. Cohen, and F. R. Volkmar, "Speech and Prosody Characteristics of Adolescents and Adults with High-Functioning Autism and Asperger Syndrome," *Journal of Speech, Language, and Hearing Research*, vol. 44, pp. 1097–1115, 2001.

[13] J. J. Diehl, D. Watson, L. Bennetto, J. McDonough, and C. Gunlogson, "An Acoustic Analysis of Prosody in High-Functioning Autism," *Applied Psycholinguistics*, vol. 30, pp. 385–404, 2009.

[14] J. P. H. van Santen, E. T. Prud'hommeaux, L. M. Black, and M. Mitchell, "Computational Prosodic Markers for Autism," *Autism*, vol. 14, pp. 215–236, 2010.

[15] R. B. Grossman, L. R. Edelson, and H. Tager-Flusberg, "Emotional facial and vocal expressions during story retelling by children and adolescents with high-functioning autism," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 3, pp. 1035–1044, 2013.

[16] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "Spontaneous-Speech Acoustic-Prosodic Features of Children with Autism and the Interacting Psychologist." in *INTERSPEECH*, 2012, pp. 1043–1046.

[17] D. Bone, C.-C. Lee, T. Chaspari, M. Black, M. Williams, S. Lee, P. Levitt, and S. Narayanan, "Acoustic-prosodic, turn-taking, and language cues in child-psychologist interactions for varying social demand," in *INTERSPEECH*, 2013.

[18] L. M. Dunn and L. M. Dunn, *Peabody Picture Vocabulary Test-Revised: PPVT-R*. American Guidance Service, 1981.

[19] R. W. Woodcock, K. McGrew, and N. Mather, *Woodcock-Johnson tests of achievement*. Itasca, IL: Riverside Publishing, 2001.

[20] R. B. Grossman, "Judgments of social awkwardness from brief exposure to children with and without high-functioning autism," *Autism*, 2014.

[21] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

[22] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," 2002.

[23] F. Hönig, A. Batliner, and E. Nöth, "Does it groove or does it stumble-automatic classification of alcoholic intoxication using prosodic features." in *INTERSPEECH*, 2011, pp. 3225–3228.

[24] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[25] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," in *Proceedings of the 17th International Congress of Phonetics Sciences*, vol. 1, 2011, pp. 1846–1849.

[26] M. Duong, J. Mostow, and S. Sitaram, "Two methods for assessing oral reading prosody," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, p. 14, 2011.

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

[28] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[29] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *Journal of autism and developmental disorders*, vol. 24, no. 5, pp. 659–685, 1994.

[30] D. Bone, C.-C. Lee, and S. Narayanan, "Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 201–213, 2014.

[31] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 1015–1028, 2011.