Acoustic-Prosodic and Turn-Taking Features in Interactions with Children with Neurodevelopmental Disorders

Daniel Bone[1]*, Somer Bishop*[2]*, Rahul Gupta*[1]*, Sungbok Lee*[1]*, Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA, USA
[2]Department of Psychiatry, UCSF School of Medicine, San Francisco, CA, USA

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

Abstract

Atypical speech prosody is a hallmark feature of autism spectrum disorder (ASD) that presents across the lifespan, but is difficult to reliably characterize qualitatively. Given the great heterogeneity of symptoms in ASD, an acoustic-based objective measure would be vital for clinical assessment and interventions. In this study, we investigate speech features in child-psychologist conversational samples, including: segmental and suprasegmental pitch dynamics, speech rate, coordination of prosodic attributes, and turn-taking. Data consist of 95 children with ASD as well as 81 controls with non-ASD developmental disorders. We demonstrate significant predictive performance using these features as well as interpret feature correlations of both interlocutors. The most robust finding is that segmental and suprasegmental prosodic variability increases for both participants in interactions with children having higher ASD severity. Recommendations for future research towards a fully-automatic quantitative measure of speech prosody in neurodevelopmental disorders are discussed.

Index Terms: prosody, autism spectrum disorder, intonation, interaction

1. Introduction

"It's not what you say, but how you say it." This common saying elucidates how critical speech prosody—the melody and timing of speech—is to effectively communicating affect and intention. Unfortunately, many verbal individuals with autism spectrum disorder (ASD) have deficits in both discerning a speaker's intent from prosody and producing appropriate prosody [1], which are detrimental to social functioning. ASD is a highly heterogeneous neurodevelopmental disorder defined by impairments in social communication and reciprocity, as well as restricted, repetitive behavioral patterns and interests [2]. Atypical prosody in ASD is considered an understudied, high impact research area [3], particularly considering the remarkable prevalence of the disorder (1 in 68 [4]).

Speech prosody of individuals with ASD is often described as exaggerated or monotone, or slow and syllable-timed [3]. Yet, inter-rater evaluation on atypical prosody for diagnostic purposes is inconsistent [5]; moreover, precision in identifying types of atypicality is low, and little is known about the prevalence of individual deficits. Prosody research can have significant translational impact; recent promising findings have shown that visual feedback intervention based on even simple prosodic measures such as vocal intensity improves production [6]. We believe that objective computational methods can support advances in the understanding and treatment of atypical prosody.

Acoustic correlates of atypical prosody have only recently been studied, as research has centered on human perception of read or spontaneous speech. Relevant findings include atypicalities in sentential [7] and contrastive stress [8], increased pausing [9], and abnormal voice quality [10]. Increasingly, computational speech scientists are taking on the task of modeling speech prosody in ASD. Studies of basic acoustic functionals have reported increased f0 variability [11] and higher maximum f0 for ASD subjects [12]. Regarding automatic modeling, researchers have: computationally measured prosodic differences in stress production [13]; automatically assessed prosodic imitation skills [14]; and classified emotional expression [15].

Our work builds on several of our previous studies which sought acoustic correlates of "atypical prosody" [16, 17, 18, 19]. It is important to note that the construct of atypical prosody is currently not well defined. As such, we have concentrated on experiments using one of two ground truths (each with their own drawbacks): either ASD diagnosis (or symptom severity) or human perception of atypicality. Our experiments [16, 17, 18] in a sample of 29 children from the USC CARE Corpus [20] found children with increasing ASD severity spoke less, spoke slower, responded later, had more variable prosody, and had more atypical voice quality. Additionally, the psychologist's cues predicted severity, given that she must continually adjust her behavior to that of the child's throughout the interaction. But not all people with ASD have prosodic difficulties, so it is desirable to relate our acoustic measures directly to perceived *atypical* prosody, even though human agreement can be rather low. We found that speech rate and rhythm cues were highly predictive of perceived "awkwardness" [19].

In this work, we continue towards an automatic prosodic evaluation by analyzing prosodic display in a large sample of individuals with ASD as well as non-ASD developmental disorders. Additionally, we introduce a novel feature group based on the coordination of prosodic modalities, and we investigate goodness of pronunciation (GOP). One limitation of the current study is that we cannot examine voice quality given potential recording differences between sites; future investigations should focus on this critical feature group. Through this study, we aim to enhance our understanding of signal-derived speech prosody measures, which are vital to behavioral interaction analyses and the creation of automated clinical tools.

2. Methodology

In the following sections we discuss: data collection and participant demographics; acoustic-prosodic features; and data analysis and machine learning.

2.1. Data Collection and Participants

Experimental data consist of Autism Diagnostic Observation Schedule (ADOS [5]) Module 3 videos of a child interacting with a psychologist. Module 3 is intended for children who are

Table 1: *Demographic information of all subjects: mean (stdv.)*

	N	Severity	Age (yr.)	NVIQ	Female
ASD	95	7.2 (2.1)	8.8 (2.6)	97.2 (20.3)	21.0%
non-ASD	81	2.6 (2.0)	8.3 (2.5)	95.7 (17.9)	30.9%

verbally fluent, and thus speech prosody is a valid analytical target. Data consist of 95 children with autism spectrum disorder (ASD) and 81 subjects with a non-ASD developmental disorder; non-ASD subject diagnoses include attention deficit hyperactivity disorder (ADHD), language disorders, mood/anxiety disorders, and intellectual disability. Participant demographics are presented in Table 1, including: ADOS severity, non-verbal IQ, age, and gender. We control for demographic differences during later analyses. ADOS severity is a measure of symptom severity from 1-10, with 10 being most severe. Subject diagnoses are "best-estimate clinical diagnoses", and consider other factors beyond the ADOS, such as parent report.

Data were collected at two sites as part of an IRB-approved study. Video and audio quality varies between sessions and sites. As diagnosis and ADOS severity was biased by site, we did not feel confident in using voice quality or energy-based measures which were previously shown to be characteristic of ASD speech [17], but could be affected by site-specific channel differences [21]. There were a total of 9 psychologists across sites. Occasionally a second psychologist or a parent was in the room. The second psychologist's actions were attributed to the primary psychologist, while a parent's actions only affected latency calculations.

2.2. Data Transcription and Alignment

The sessions were first manually transcribed through use an adapted version of the SALT guidelines [22], wherein utterances were also manually segmented. Speech segments that were unintelligible or that contained high levels of background noise were excluded from further acoustic analysis, as were sessions that were entirely noisy. Word, syllable, and phonemic forced-alignment were then performed using data-specific acoustic models with Kaldi [23].

2.3. Acoustic-Prosodic and Turn-taking Features

We compute five classes of features: turn-taking and speaking rate; segmental pitch cues; suprasegmental intonation features; goodness of pronunciation; and coordination between prosodic modalities (a novel feature type). Details of the feature computation for each group follows in this section.

2.3.1. Turn-taking and Speaking Rate

We compute seven temporal descriptors of the social interaction based on our previous work [16]. Six turn-taking features describe the conversational style of each participant: speaking time (%), turn length (*words*), latency (*s*), overall silence time (%), intra-turn pause length (*s*), and fraction intra-turn pausing (%). Speaking rate is calculated using forced-aligned syllable boundaries as ($\#syllables/s$). All features are calculated as the median over a session.

2.3.2. Segmental Pitch Cues: Syllabic Pitch Contours

We consider the segmental intonation contours of short lexical units as in previous studies [21, 17, 19]. This technique may capture speaker idiosyncrasies in micro-prosodic production. We calculate syllable-level second-order polynomial parametrization of pitch, then calculate session-level medians and inter-quartile ratios of slope and curvature. The overall median pitch is also calculated, totaling 5 features.

2.3.3. Suprasegmental Intonation: Momel/Intsint

We characterize individuals' intonation patterns in order to quantify perceptions of either monotonic or exaggerated intonation. In particular, the macro-prosodic movement of pitch is modeled using an automatic signal-derived method, Momel (MOdeling MELody), which provides a phonetic representation of pitch intonation patterns [24]. The algorithm produces a smooth curve that models the macro-melodic movements of pitch, where deviations are attributed to micro-prosodic movements related to segmental constraints. Taking raw fundamental frequency as input, a set of target points for quadratic interpolation is output. We compute the median-absolute-difference between Momel points to capture dynamic variability.

These target points can be further transformed into a symbolic representation of fundamental frequency patterns, namely Intsint (International Transcription System for Intonation [24]). Intsint comprises a limited set of abstract tonal symbols, grouped as absolute or relative. Absolute tones refer to a speaker's overall pitch range, and are categorized as top, mid, or bottom (T, M, B). Relative tones are determined relative to the previous tone, and are categorized as: same (S) if less than a threshold from the previous target; non-iterative high/low step (H/L); or iterative up/down step (U/D), which tend to be smaller than H or L steps. An example intonation contour and corresponding Momel and Intsint outputs is displayed in Figure 1.

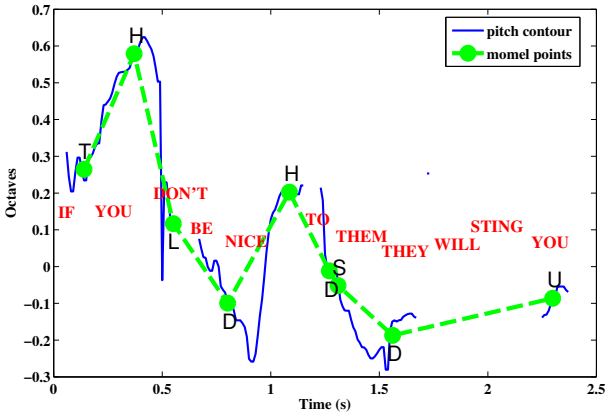

Figure 1: *Example intonation contour plotted with Momel target points and Intsint symbolic representation.*

We calculate the frequency of relative tone changes (H, U, S, D, L) as features, which may capture a speaker's tendency towards specific pitch dynamics. We implemented versions of the proposed Momel/Intsint algorithms [25] in Matlab. Thresholds for small and large steps were set at 0.125 and 0.25 octaves from center. Analysis is done in the OME (Octave MEdian) scale [26], a log-pitch transformation as in Eq. 1 through which speaker's tend to have the same pitch range of one octave.

$$\text{OME} = \log(f0_{Hz}) - \log(\text{median}(f0_{Hz})) \quad (1)$$

Since a speaker's range has been observed to reliably be one OME around center in neutral speech, all speakers should have a comparable range regardless of median pitch (unlike for Hz).

2.3.4. Goodness of Pronunciation

Phonemic spectral distortions due to atypical, or immature, articulatory controls may be perceived as "atypical" speech production. As such, we utilize a measure of phonetic pronunciation quality, goodness of pronunciation (GOP [27]). GOP has been shown useful for other paralinguistic tasks such as nativeness detection [28]. GOP uses acoustic models (AMs) trained

Table 2: *Correlations of features with ADOS severity and best-estimate diagnosis. * indicates $p<0.05$; n.s. is non-significant.*

Category	Feature	Child			Psychologist		
		Trend with severity	*Sp. ρ*	*group diff.?*	*Trend with severity*	*Sp. ρ*	*group diff.?*
Turn-taking & speaking rate	speaking time (%)	less	-0.15^*		*n.s.*	0.01	
	turn length (words)	shorter	-0.17^*		*n.s.*	0.13	
	intra-turn pause (s)	longer	0.18^*	✓	*n.s.*	0.12	
	intra-turn pause (%)	*n.s.*	0.00		more	0.17^*	✓
	latency (sec)	*n.s.*	0.11		longer	0.24^*	
	silence (%)	more	0.15^*		more	0.15^*	
	speaking rate (syl/s)	slower	-0.20^*	✓	*n.s.*	-0.05	
Segmental pitch cues	f0 curve median	lower	-0.18^*		*n.s.*	-0.09	
	f0 slope median	*n.s.*	0.12		*n.s.*	0.12	
	f0 curve IQR	more	0.36^*	✓	*n.s.*	0.08	
	f0 slopeIQR	more	0.28^*		more	0.25^*	✓
	f0 median	higher	0.25^*	✓	higher	0.21^*	
Suprasegmental Intonation	m.a.d. Momel	higher	0.17^*	✓	higher	0.25^*	✓
	Intsint *High* Tone (%)	higher	0.19^*		higher	0.32^*	✓
	Intsint *Same* Tone (%)	lower	-0.19^*		lower	-0.24^*	✓
	Intsint *Low* Tone (%)	higher	0.19^*		higher	0.31^*	✓
Prosodic Coordination	corr. f0 & dur.	lower	-0.25^*	✓	*N/A*	*N/A*	*N/A*
	corr. f0 & intensity	lower	-0.17^*		*N/A*	*N/A*	*N/A*
	corr. dur. & intensity	*n.s.*	0.09		*N/A*	*N/A*	*N/A*
Pronunciation	GOP	lower	-0.20^*		*N/A*	*N/A*	*N/A*

on domain data to quantify pronunciation quality:

$$\text{GOP} = \frac{1}{N} \log\left(\frac{P(O|\text{Transcript})}{P(O|\text{AM loop})}\right) \quad (2)$$

where N is the number of frames and O are the acoustic features. The numerator is the likelihood of the acoustics, given the transcription and the native AMs, which is equivalent to forced alignment. The denominator is typically estimated via automatic speech recognition with the same set of AMs and an "AM loop" grammar. Higher GOP scores indicate higher pronunciation quality. GOP computation is performed using Kaldi [23] with a slight modification. In our implementation of computing the denominator, we do not allow for transitions between phones within the forced aligned phone boundaries.

2.3.5. Prosodic Coordination Features

We suspected that individuals with ASD may coordinate prosodic modalities in unique ways. To assess this hypothesis, we introduce a feature which measures the simultaneous movements of pitch, duration, and intensity. In particular, for each syllable we compute the duration, median pitch, and median intensity. These features are concatenated per speaker, and then the Spearman's rank-correlation coefficient is calculate pairwise, producing three features.

2.4. Statistical Analysis and Machine Learning

We conduct correlation analysis, as well as classification (support vector machine) and regression (support vector regression, SVR) via Liblinear software [29]. Parameters are tuned using two-level nested cross-validation, and average statistics of ten runs of CV are reported. Spearman's rank-correlation coefficient and unweighted average recall (UAR, the mean of per-class recall) are selected for evaluation metrics.

3. Results and Discussion

In this section we explore prosodic variation within ASD based on objective features. In Section 3.1, the objective cues are analyzed in the context of interaction, and in Section 3.2, the cues are used to predict ASD severity (from the ADOS) and best-estimate clinical diagnosis (ASD, non-ASD).

3.1. Correlational Feature Analysis

Acoustic and turn-taking feature correlations for both child and psychologist are provided in Table 2. We concentrate on correlations with ASD severity, which is better explained by our features than best-estimate diagnosis. This finding may stem from the fact that ASD severity is calculated from the ADOS interaction data, whereas best-estimate diagnosis draws from external factors that we cannot observe. All significant correlations with severity are still significant after controlling for demographic variables (i.e., age, gender, and NVIQ) except child turn length (p=0.07) and child pitch curvature median (p=0.11).

Turn-taking cues provide an overarching depiction of interaction quality. In our data, children with higher ASD severity tend to speak less, in shorter phrases, and with longer pauses; additionally, they speak slower on average. These findings mirror our previous findings in a smaller dataset[18]. Since the psychologist is not only the evaluator, but also a participant in the interaction, we can observe effects in their behavior according to their participant's social cues. In particular, the psychologist tends to pause more within a turn and wait longer to start a turn (more latency). Overall, there is also more silence. In sum, this suggests that the psychologist may be unsure of when the child will start and end turns, or that the child may be unresponsive.

Segmental pitch cues show short-term variability in use of fundamental frequency. Children with higher social-communicative deficits showed negative pitch curvature, which is possibly perceived as "flat" or "monotonic". For both the child and the psychologist, short-term dynamic variability of fundamental frequency increases. Pitch variability may increase with enhanced affect, such as the psychologist trying to engage the child, or in response to an interlocutor's behavioral patterns.

Suprasegmental cues are essential for communicating intention and affect. While we cannot fully model prosody without knowing the semantic context of an uttered phrase, we can look at global tendencies. Speakers with higher severity are shown to have more macro-prosodic variability in all four of the features that we examine, and likewise for the psychologist. Specifically, both participants have larger pitch movements (octaves) between successive Momel points. In symbolic repre-

sentation (Intsint), there are more *high* and *low* tones, and less *same* level pitch movements. This finding expands on previous reports of higher pitch variability, which often did not use log-scales (pitch is log-normally distributed within speaker) and were simply global functionals on raw fundamental frequency, not providing insight into the dynamics. Note that the intermediate *up-step* and *down-step* tones are not displayed to improve readability, since neither reached significance.

After listening to speaking samples, we suspected that individuals with "atypical" prosody were sometimes modulating a prosodic modality independent of other modalities. We quantified prosodic coordination as the pairwise coordination between three modalities: syllabic fundamental frequency, vocal intensity, and duration. Results support that children with higher ASD symptom severity coordinate their use of pitch with duration and vocal intensity to a lesser extent.

Lastly, we investigate pronunciation quality, motivated by the possibility that articulation distortions, which occur generally in those with language delays, may be attributed to atypical prosody. Results show that children with higher ADOS-ASD severity do tend to have a lower goodness of pronunciation. Whether, and to what degree, articulation distortions affect perceptions of atypical prosody is a topic of future research.

3.2. Prediction Experiments

Correlational analysis informs interpretation in behavioral interactions, but computational systems that support clinical researchers in behavior tracking and intervention can rely on joint modeling of many features. In this section, we analyze the performance of different feature categories in predicting ASD severity and best-estimate diagnosis (Table 3).

We initially examine the predictive power of the baseline demographic features: age, gender, and non-verbal IQ. If these features are predictive on their own, it's possible that our speech cues are directly predictive of demographics (e.g., IQ or age), rather than ASD-related social behavior. We find that the demographic features are not significantly predictive of ASD severity or diagnosis, and thus conclude that our features are capturing ASD-specific behavioral patterns.

All features groups are significantly predictive of ASD severity, and all but goodness of pronunciation significantly classify ASD from non-ASD interactions based on both the child's and the psychologist's features. Pronunciation quality may more directly affect or represent social functioning, since there is no diagnostic relevance.

The top individual feature groups for predicting severity are: the child's segmental pitch features; the psychologist's suprasegmental intonation features; and the child's and psychologist's combined segmental intonation features. Based on the analysis of Section 3.1, we conclude that increased prosodic variability of both participants is a reliable predictor of ASD severity. The individual feature groups achieve similar ASD/non-ASD classification performance (aside from pronunciation quality); turn-taking and speaking rate statistics reach a peak of 58% unweighted average recall. Feature fusion leads to the optimal performance of 0.35 correlation and 59% UAR.

Table 3: *Regression and classification of ASD severity and best-estimate diagnosis via acoustic-prosodic and turn-taking features.* **Bolded** *statistics are significant at the* α=0.05 *level.*

	ASD Severity			**Diagnosis**
Features	*Child*	*Psych.*	*C.&P.*	*C.&P.*
Baseline: Demographics	-0.02	*N/A*	*-0.02*	52%
Turn-taking	**0.17**	**0.17**	**0.18**	**58%**
Segmental	**0.28**	**0.19**	**0.30**	**57%**
Suprasegmental	0.08	**0.25**	**0.22**	**56%**
Prosodic Coord.	**0.17**	*N/A*	**0.17**	**56%**
Pronunciation	**0.17**	*N/A*	**0.17**	52%
Feature Fusion	**0.31**	**0.31**	**0.35**	**59%**
metric	Spearman's ρ_S			UAR

4. Conclusion

We examined acoustic-prosodic and turn-taking features in interactions with individuals with neurodevelopmental disorders towards a better, evidence-based understanding. Five groups of features were considered: turn-taking and speaking rate, suprasegmental intonation, segmental pitch, prosodic coordination, and pronunciation quality. Unfortunately, voice quality features were excluded due to potential channel differences between sites. The most robust finding is that segmental and suprasegmental prosodic variability increases for both participants in interactions with children having higher ASD severity (or ASD versus non-ASD disorders). Additionally, based on our proposed features, children with higher ASD severity showed lower coordination of pitch with other modalities.

5. Outlook for Future Research

Speech prosody remains a critical research area in autism spectrum disorder for which objective assessment can have true impact in characterizing and tracking prosodic deficits. However, one of the primary reasons that speech prosody is understudied in autism is because of the difficulty in modeling it during conversational speech due to its variability and dynamic nature. Initial studies were limited to features like the mean and standard deviation of pitch and intensity. The following are suggestions for future research towards the goal of *creating a computational characterization of prosody in neurodevelopmental disorders.*

- **Optimal data collection:** Collected should have high quality (for complex feature extraction), high consistency, and ecological validity. Spontaneous speech is much preferred over read due to relevance to this social-communicative disorder.
- **Maintaining Interpretability:** A great appeal of engineering methods is that complex models can be created that humans need not understand, e.g., deep learning. However, in this particular problem domain, the primary drawback is that interpretability is largely abandoned. With a loss of interpretability, it is difficult to track why a system is successful (which may be for dubious reasons [21]), and it is less clear if the system will generalize to independent, uniquely collected data (than, for example, knowledge-drive approaches [30]).
- **Selecting a Ground Truth:** Supervised learning necessitates a ground truth. However, atypical prosody is a construct that, while critically important, has no reliable ground truth. Two choices are apparent: either ASD/non-ASD diagnosis or human judgment. ASD behavior is highly variable; since not all children have the same deficits, ASD diagnosis cannot be equated to "autistic" prosody. Alternatively, human judgment is unreliable, especially for untrained raters [19]. Thus, it is our suggestion that future studies simultaneously analyze the relevance of prosodic features against both ground truths. Moreover, it may be necessary that the final objective measures are entirely bottom-up, derived from and defined by signals. Such a rule-based approach would come with its own difficulties in generalization, but would be one solution to creating a fully objective definition of atypical prosody.

6. Acknowledgments

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

7. References

[1] 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.

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

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

[4] J. Baio, "Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010." *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)*, vol. 63, no. 2, p. 1, 2014.

[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] E. S. Simmons, R. Paul, and F. Shic, "Brief report: A mobile application to treat prosodic deficits in autism spectrum disorder and other communication impairments: A pilot study," *Journal of autism and developmental disorders*, vol. 46, no. 1, pp. 320–327, 2016.

[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] 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.

[12] 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.

[13] 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.

[14] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for the prosodic assessment of language-impaired children," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1328–1342, 2011.

[15] E. Marchi, B. Schuller, S. Baron-Cohen, O. Golan, S. Bölte, P. Arora, and R. Häb-Umbach, "Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[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] ——, "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.

[18] 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.

[19] D. Bone, M. P. Black, A. Ramakrishna, R. Grossman, and S. Narayanan, "Acoustic-prosodic correlates of awkwardprosody in story retellings from adolescents with autism," 2015.

[20] M. P. Black, D. Bone, M. E. Williams, P. Gorrindo, P. Levitt, and S. S. Narayanan, "The USC CARE Corpus: Child-Psychologist Interactions of Children with Autism Spectrum Disorders," in *Proceedings of Interspeech*, 2011.

[21] 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.

[22] J. Miller and R. Chapman, "Systematic analysis of language transcripts," *Madison, WI: Language Analysis Laboratory*, 1985.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[24] D. Hirst, A. Di Cristo, and R. Espesser, "Levels of representation and levels of analysis for the description of intonation systems," in *Prosody: Theory and experiment*. Springer, 2000, pp. 51–87.

[25] D. Hirst, "A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation," in *Proceedings of the XVIth International Conference of Phonetic Sciences*, vol. 12331236, 2007.

[26] C. De Looze and D. Hirst, "The ome (octave-median) scale: A natural scale for speech prosody," in *Proceedings of the 7th International Conference on Speech Prosody (SP7)*, 2014.

[27] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[28] M. P. Black, D. Bone, Z. I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S. N. Chakravarthula, B. Xiao, M. Van Segbroeck, J. Kim *et al.*, "Automated evaluation of non-native english pronunciation quality: Combining knowledge-and data-driven features at multiple time scales," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[29] 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.

[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.