NIH Public Access
Author Manuscript
J Speech Lang Hear Res. Author manuscript; available in PMC 2015 February 12.

Published in final edited form as:
J Speech Lang Hear Res. 2014 August ; 57(4): 1162–1177. doi:10.1044/2014_JSLHR-S-13-0062.

The Psychologist as an Interlocutor in Autism Spectrum Disorder Assessment: Insights From a Study of Spontaneous Prosody

Daniel Bone[a], **Chi-Chun Lee**[a], **Matthew P. Black**[a], **Marian E. Williams**[b], **Sungbok Lee**[a], **Pat Levitt**[c,d], and **Shrikanth Narayanan**[a]

[a]Signal Analysis & Interpretation Laboratory (SAIL), University of Southern California, Los Angeles

[b]University Center for Excellence in Developmental Disabilities, Keck School of Medicine of University of Southern California and Children's Hospital Los Angeles

[c]Keck School of Medicine of University of Southern California

[d]Children's Hospital Los Angeles

Abstract

Purpose—The purpose of this study was to examine relationships between prosodic speech cues and autism spectrum disorder (ASD) severity, hypothesizing a mutually interactive relationship between the speech characteristics of the psychologist and the child. The authors objectively quantified acoustic-prosodic cues of the psychologist and of the child with ASD during spontaneous interaction, establishing a methodology for future large-sample analysis.

Method—Speech acoustic-prosodic features were semiautomatically derived from segments of semistructured interviews (Autism Diagnostic Observation Schedule, ADOS; Lord, Rutter, DiLavore, & Risi, 1999; Lord et al., 2012) with 28 children who had previously been diagnosed with ASD. Prosody was quantified in terms of intonation, volume, rate, and voice quality. Research hypotheses were tested via correlation as well as hierarchical and predictive regression between ADOS severity and prosodic cues.

Results—Automatically extracted speech features demonstrated prosodic characteristics of dyadic interactions. As rated ASD severity increased, both the psychologist and the child demonstrated effects for turn-end pitch slope, and both spoke with atypical voice quality. The psychologist's acoustic cues predicted the child's symptom severity better than did the child's acoustic cues.

Conclusion—The psychologist, acting as evaluator and interlocutor, was shown to adjust his or her behavior in predictable ways based on the child's social-communicative impairments. The results support future study of speech prosody of both interaction partners during spontaneous

© American Speech-Language-Hearing Association

Correspondence to Daniel Bone: dbone@usc.edu.

Disclosure: The authors have declared that no competing interests existed at the time of publication.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

conversation, while using automatic computational methods that allow for scalable analysis on much larger corpora.

Keywords

autism spectrum disorder; children; prosody; social communication; assessment; dyadic interaction

Human social interaction necessitates that each participant continually perceive, plan, and express multimodal pragmatic and affective cues. Thus, a person's ability to interact effectively may be compromised when there is an interruption in any facet of this perception–production loop. *Autism spectrum disorder (ASD)* is a developmental disorder defined clinically by impaired social reciprocity and communication—jointly referred to as *social affect* (Gotham, Risi, Pickles, & Lord, 2007)—as well as by restricted, repetitive behaviors and interests (American Psychiatric Association, 2000).

Speech *prosody*—which refers to the manner in which a person utters a phrase to convey affect, mark a communicative act, or disambiguate meaning—plays a critical role in social reciprocity. A central role of prosody is to enhance communication of intent and, thus, enhance conversational quality and flow. For example, a rising intonation can indicate a request for response, whereas a falling intonation can indicate finality (Cruttenden, 1997). Prosody can also be used to indicate affect (Juslin & Scherer, 2005) or attitude (Uldall, 1960). Furthermore, speech prosody has been associated with social-communicative behaviors such as eye contact in children (Furrow, 1984).

Atypical prosody has been regularly reported in individuals with ASD. Furthermore, atypical prosody is relevant to certain overarching theories on ASD—for example, impaired theory of mind (Baron-Cohen, 1988; Frith, 2001; Frith & Happé, 2005; McCann & Peppe, 2003). Specifically, inability to gauge the mental state of an interlocutor may be due to impairments in perception of prosody, which in turn may create challenges for producing appropriate prosodic functions. Many studies have investigated receptive and expressive language skills in autism (e.g., Boucher, Andrianopoulos, Velleman, Keller, & Pecora, 2011; Paul, Augustyn, Klin, & Volkmar, 2005). Tested theories include the *speech attunement framework* (Shriberg, Paul, Black, & van Santen, 2011)—which decomposes production–perception processes into "tuning in" to learn from the environment and "tuning up" one's own behavior to a level of social appropriateness—as well as disrupted speech planning and atypical motor system function such as that seen in childhood apraxia of speech (American Speech-Language-Hearing Association, 2007a, 2007b). Given the complexity of developing speech, it is not surprising that the mechanisms through which atypical prosody occurs in children with ASD remain unclear.

Atypical Prosody in ASD

Qualitative descriptions of prosodic abnormalities appear throughout the ASD literature, but contradictory findings are common, and the specific features of prosody measured are not always well defined (McCann & Peppe, 2003), a testament to both their relevance and the challenges in standardizing prosodic assessment. For example, pitch range has been reported

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

as both exaggerated and monotone in individuals with ASD (Baltaxe, Simmons, & Zee, 1984). Characterization of prosody is also incorporated within the widely used diagnostic instruments, the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 1999, 2012) and the Autism Diagnostic Interview—Revised (ADI–R; Rutter, LeCouteur, & Lord, 2003). The ADOS considers any of the following qualities to be characteristic of speech associated with ASD: "slow and halting; inappropriately rapid; jerky and irregular in rhythm … odd intonation or inappropriate pitch and stress, markedly flat and toneless … consistently abnormal volume" (Lord et al., 1999, Module 3, p. 6), and the ADI–R prosody item focuses on the parent's report of unusual characteristics of the child's speech, with specific probes regarding volume, rate, rhythm, intonation, and pitch. A variety of markers can contribute to a perceived oddness in prosody such as differences in pitch slope (Paccia & Curcio, 1982), atypical voice quality (Sheinkopf, Mundy, Oller, & Steffens, 2000), and nasality (Shriberg et al., 2001). This inherent variability and subjectivity in characterizing prosodic abnormalities poses measurement challenges.

Researchers have used structured laboratory tasks to assess prosodic function more precisely in children with ASD. Such studies have shown, for instance, that both sentential stress (Paul, Shriberg, et al., 2005) and contrastive stress (Peppe, McCann, Gibbon, O'Hare, & Rutherford, 2007) differed in children with ASD compared with typical peers. Peppe et al. (2007) developed a structured prosodic screening profile that requires individuals to respond to computerized prompts; observers rate the expressive prosody responses for accuracy in terms of delivering meaning. However, as Peppe (2011) remarked, the instrument "provides no information about aspects of prosody that do not affect communication function in a concrete way, but may have an impact on social functioning or listenability … such as speech-rhythm, pitch-range, loudness and speech-rate" (p. 18).

In order to assess these global aspects of prosody that are thought to differ in individuals with atypical social functioning, researchers have used qualitative tools to evaluate prosody along dimensions such as phrasing, rate, stress, loudness, pitch, laryngeal quality, and resonance (Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997; Shriberg et al., 2001, 2010). Although these methods incorporate acoustic analysis with software in addition to human perception, intricate human annotation is still necessary. Methods that rely on human perception and annotation of each participant's data are time intensive, limiting the number of participants that can be efficiently studied. Human annotation is also prone to reliability issues, with marginal to inadequate reliability found for item-level scoring of certain prosody voice codes (Shriberg et al., 2001). Therefore, automatic computational analysis of prosody has the potential to be an objective alternative or complement to human annotation that is scalable to large data sets—an appealing proposition given the wealth of spontaneous interaction data already collected by autism researchers.

Transactional Interactions and ASD

In addition to increased understanding of the prosody of children with autism, this study paradigm allows careful examination of prosodic features of the psychologist as a communicative partner interacting with the child. Synchronous interactions between parents and children with ASD have been found to predict better long-term outcomes (Siller &

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Sigman, 2002), and many intervention approaches include an element of altering the adult's interactions with the child with ASD to encourage engaged, synchronous interactions. For example, in the social communication, emotional regulation, and transactional support (SCERTS) model, parents and other communication partners are taught strategies to "attune affectively and calibrate their emotional tone to that of the less able partner" (Prizant, Wetherby, Rubin, & Laurent, 2003, p. 308). Changes in affective communication and synchrony of the caregiver or interventionist with the child are also elements used in pivotal response training (e.g., Vernon, Koegel, Dauterman, & Stolen, 2012), DIR/Floortime (e.g., Weider & Greenspan, 2003), and the Early Start Denver Model (Dawson et al., 2010). The behavior of one person in a dyadic interaction generally depends intricately on the other person's behavior—evidenced in the context provided by age, gender, social status, and culture of the participants (Knapp & Hall, 2009) or the behavioral synchrony that occurs naturally and spontaneously in human–human interactions (Kimura & Daibo, 2006). Thus, we investigated the psychologist's acoustic-prosodic cues in an effort to understand the degree to which the interlocutor's speech behavior varies based on interaction with participants of varying social-affective abilities.

Current Study Goals and Rationale

Because precise characterization of the global aspects of prosody for ASD has not been established (Diehl, Watson, Bennetto, McDonough, & Gunlogson, 2009; Peppe et al., 2007), the current study presents a strategy to obtain a more objective representation of speech prosody through signal processing methods that quantify qualitative perceptions. This approach is in contrast to experimental paradigms of constrained speaking tasks with manual annotation and evaluation of prosody by human coders (Paul, Shriberg, et al., 2005; Peppe et al., 2007). Furthermore, previous studies have been limited primarily to the analysis of speech of children with high-functioning autism (HFA) out of the context in which it was produced (Ploog, Banerjee, & Brooks, 2009). Although clinical heterogeneity may explain some conflicting reports regarding prosody in the literature, analysis of more natural prosody through acoustic measures of spontaneous speech in interactive communication settings has the potential to contribute to better characterization of prosody in children with ASD.

The present study analyzed speech segments from spontaneous interactions between a child and a psychologist that were recorded during standardized observational assessment of autism symptoms using the ADOS. The portions of the assessment that were examined represent spontaneous interaction that is constrained by the introspective topics and interview style. Spontaneous speech during the ADOS assessment has been shown to be valid for prosodic analysis (Shriberg et al., 2001).

Prosody is characterized in terms of the global dynamics of intonation, volume, rate, and voice quality. Regarding potential acoustically derived correlates of perceived abnormalities in these speech segments, few studies offer suggestions (Diehl et al., 2009; van Santen, Prud'hommeaux, Black, & Mitchell, 2010), and even fewer have additionally assessed spontaneous speech (Shriberg et al., 2011). As such, the current study proposes a set of acoustic-prosodic features to represent prosody in child–psychologist dialogue.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

A crucial aim of this work was to incorporate analysis of the acoustic-prosodic characteristics of a psychologist engaged in ADOS administration rather than to focus only on the child's speech. This transactional, dyadic focus provides an opportunity to discern the adaptive behavior of the psychologist in the context of eliciting desired responses from each child and to examine possible prosodic attunement between the two participants.

Some researchers have called for a push toward both dimensional descriptions of behavior and more valid and reliable ways to quantify such behavior dimensions (e.g., Lord & Jones, 2012). This work—part of the emerging field of *behavioral signal processing* (BSP; Narayanan & Georgiou, 2013)—attempts to address these goals. For instance, such computational approaches have lent quantitative insight into processes such as prosodic entrainment between interacting dyads and affectivity patterns (Lee et al., 2014). The co-variation between continuous behavioral descriptors of speech prosody and dimensional ratings of social-affective behavior is investigated in the present article. Given the apparent continuum of phenotypic behavior, correlational analysis using ordinal-scale behavior ratings may prove invaluable toward effective stratification that supports further study (e.g., genetic research).

This article provides a more detailed analysis than was documented in a previous report on spontaneous prosody during the ADOS (Bone, Black, et al., 2012). The overarching goal is to develop a framework for large-sample analysis of prosody, in a dyadic setting, by using semiautomatic computational methods. The validation of the strategy to perform large-scale analysis of natural speech data between clinician and child has the potential to provide greater insight for developing more effective ASD interventions. The specific aims addressed in the present study include (a) demonstration of the feasibility of semiautomatic computational analysis of specific, perceptually inspired acoustic-prosodic elements of speech during naturalistic conversational interchange in children with ASD; (b) exploration of the relationship between prosodic features in the speech of children with ASD and those of the psychologist interlocutors; (c) exploration of the relationship between children's autism symptom severity and the prosodic features of their speech; and (d) exploration of the relationship between children's autism symptom severity and the prosodic features of the psychologist during interaction with each child. We hypothesized that the psychologist's prosody and the child's prosody would vary depending on the level of severity of ASD symptoms of each child.

Method

The research design and method was approved by the institutional review boards of Children's Hospital Los Angeles and the University of Southern California, and written informed consent was obtained from the parents of all participants. Exclusion criteria included severe sensory or motor impairment, neurodevelopmental disorders of known etiology (Rett syndrome, tuberous sclerosis, Down syndrome, phenylketonuria, 22Q deletion syndrome, fragile X syndrome, and neurofibromatosis), gestational age of less than 36 or greater than 42 weeks, and birth weight less than 2,500 g.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Participants

Participants were recruited as part of a larger study of children with ASD, with or without co-occurring medical conditions. The present study included 28 children without a diagnosed or parent-reported medical condition, ranging in age from 5.8 to 14.7 years (*M* = 9.8, *SD* = 2.5). Of the 28 participants, 22 (79%) were male, six (21%) were female, 20 (71%) were Hispanic, and eight (29%) were White, Non-Hispanic. Parents were asked to report the child's primary or first language. The first languages of the 28 participants were English (15 children, 54%), Spanish (nine children, 32%), and both English and Spanish (four children, 14%).

These data are a subset of the USC Center for Autism Research in Engineering (CARE) Corpus (Black et al., 2011). The behavioral data were collected as a part of a larger genetic study for which the ADOS was administered to confirm the ASD diagnosis. Age for inclusion was 5–17 years, and for this sample, prior diagnosis of an autism spectrum disorder by a professional in the community was required. All verbally fluent children from the larger study were included in this sample, determined on the basis of the psychologist's decision to administer Module 3 of the ADOS (see the first subsection in the Measures section below).

Confirmation of autism diagnosis was established by the psychologist on the basis of ADOS scores, any input provided by the parent during the assessment, and review of available records of the previous diagnosis. In this sample, 17 (61%) of the participants had a confirmed diagnosis of autism on the ADOS, five (18%) had a diagnosis of ASD but not full autism, and six (21%) scored below the cutoff for ASD on the ADOS—meaning that they were deemed to not have ASD.

Children whose parent(s) spoke primarily Spanish were assessed by a bilingual (Spanish/English) psychologist, and children had the option to respond in Spanish or to request Spanish interactions if they felt more comfortable conversing in Spanish. This sample includes only children who chose to participate in the assessment in English; one participant was excluded from this analysis due to a primarily Spanish discourse. Another participant was excluded due to nominal vocal activity (verbal or nonverbal) during the assessment, which furthermore was muffled and unintelligible.

In addition to speech data from children, this study includes speech data from the three licensed psychologists who administered the ADOS for the genetic study. All three psychologists were women, and all were research-certified in the ADOS and had extensive clinical experience working with children with ASD. Two psychologists were bilingual in English and Spanish; one was a native Spanish speaker who was also fluent in English.

Measures

ADOS—The ADOS was administered by one of three psychologists with research certification in the measure. The ADOS is a standardized assessment of autism symptoms conducted through a series of activities designed to elicit a sample of communication, social interaction, play, and other behaviors. The ADOS is designed with different modules, chosen based primarily on the child's level of expressive language. The present study

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

includes participants who were administered only Module 3, designed for children with *fluent speech*, defined according to the ADOS manual as speech that includes “a range of flexible sentence types, providing language beyond the immediate context, and describing logical connections within a sentence” (Lord et al., 1999, p. 5). In order to identify the child’s level of verbal fluency, the administering psychologist followed a three-step process. First, the parent answered a series of questions about the child’s language level by telephone prior to the session. Next, the psychologist interacted with the child in the clinic while the research assistant was obtaining informed consent to further confirm the child’s level of verbal fluency. If the child spoke in complete utterances during this interaction, the psychologist proceeded with administering Module 3. The psychologist then continued to assess the child’s verbal fluency during the first 10 min of the ADOS session. Following the standard ADOS protocol, the psychologist changed modules after the first part of the assessment if the child’s expressive language did not fit the definition of fluent speech in the ADOS manual required for Module 3. For this study, only participants who were administered Module 3 were included. Formal language assessment was not conducted as part of the larger study, so data about the relative language skills of the participants could not be presented.

All ADOS evaluations were audio and video recorded. The evaluations took place in a single, multi-use clinical room. A portable recording setup was used, with all sensors operating in the far-field to maintain diagnostic validity. *Far-field* refers to the extended distance of the sensors to the target; in our case, the high-quality microphones and cameras were roughly 2 m from the participants. Two Sony HDR-SR12 High-Definition Handycam camcorders were mounted on tripods in two corners of the room. Additional audio recordings were collected from two high-quality directional shotgun microphones (SCHOEPS CMIT 5 U), which were mounted next to the camcorders. The uncompressed audio was captured with an Edirol R-4 Pro recorder (48 kHz, 24 bit). This study analyzed down-sampled audio (16 kHz) from a single channel of one high-quality microphone, chosen on the basis of perceived quality of the recordings.

Targeted ADOS Activities—The speech samples for the present study were obtained from two of the standard ADOS activities: (a) Emotions and (b) Social Difficulties and Annoyance. These activities were selected because each offers a continuous sampling of conversational speech, rich with emotionally focused content pertinent to ASD diagnosis. A child with ASD may be less comfortable communicating about these particular topics than their typically developing peers, which should be noted during interpretation of results. Because the conversational style of these two subtasks is rather constrained, such apprehension may be implicitly captured by the automatic measures. From the start of the first selected activity (usually Emotions, although, during standard ADOS administration, the assessor can change the order of administration of activities to maintain rapport), we collected up to 5 min per session for analysis (minimum = 101 s, *M* = 264 s, *SD* = 8.4 s). Because there is variability in the duration of analyzed data across subjects, all extracted speech features were designed to be independent of the duration of the data (i.e., robust statistics, such as medians and interquartile ratios, were used).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

ASD Severity—ADOS Module 3 includes 28 codes scored by the examiner immediately following the assessment. The *diagnostic algorithm* consists of a subset of the codes used to determine if the child's scores exceed the cutoffs typical of children with autism in the standardization group for the measure. For this analysis, we used the revised algorithms (Gotham et al., 2007) rather than the original ADOS algorithm because the revised algorithms are based on more extensive research regarding the codes that best differentiate children with ASD from typically developing children. Algorithm scores were then converted to an autism symptom severity score, following the recommendation of Gotham, Pickles, and Lord (2009). The dependent variable in this study was the severity score, which is based on the Social Affect and Restricted, Repetitive Behaviors factors in the revised ADOS diagnostic algorithm and the severity scale that is used for normalization across modules and age (Gotham et al., 2009).

ADOS severity was analyzed instead of the atypical prosody ADOS code, Speech Abnormalities Associated With Autism, for three reasons: (a) Atypical prosody is difficult to describe and relies on subjective interpretation of multiple factors; (b) atypical prosody in the ADOS is coded on a low-resolution three-point scale; and (c) the atypical prosody ADOS code is highly correlated with overall ADOS severity—in our data set of interest, $r_s(26) = 0.73$, $p < .001$.[1]

Prosodic Quantification—A primary goal of this study was to capture disordered prosody by direct speech-signal-processing techniques in such a way that it may scale more readily than full-hand annotation. Twenty-four features (number of each type denoted parenthetically) were extracted that address four key areas of prosody relevant to ASD: pitch (6), volume (6), rate (4), and voice quality (8). These vocal features were designed through referencing linguistic and engineering perceptual studies in order to capture the qualitatively described disordered prosody reported in the ASD literature. The features are detailed in the subsections that follow. In order to determine whether meaningful variations in the psychologist's voice corresponded to the child's behaviors, we also extracted the same prosodic features from the psychologist's speech. The signal analysis used here can be considered semiautomatic because it takes advantage of manually derived text transcripts for accurate automatic alignment of the text to the audio, as described next.

Text-to-speech alignment: A necessary objective of this study was to appropriately model the interaction with meaningful vocal features for each participant. For many of the acoustic parameters that we extracted, it was necessary to understand when each token (word or phoneme) was uttered within the acoustic waveform. For example, detecting the start and end times of words allows for the calculation of syllabic speaking rate, and the detection of vowel regions allows for the computation of voice quality measures. Manual transcription at this fine level is not practical or scalable for such a large corpus; thus, we relied on computer speech-processing technologies. Because a lexical-level transcription was available with the

[1] This correlation was also calculated on the much larger, distinct Autism Genetic Resource Exchange (AGRE; Geschwind et al., 2001) database and was again found to be significant, but with medium effect size, $r_S(1139) = 0.48$, $p < .001$. The AGRE Module 3 phenotypic data that we used were downloaded on April 6, 2013. The data comprised 1,143 subjects with a mean age of 9.5 years (σ= 3.0 years). Two of the 1,143 subjects were excluded for missing ADOS code data, leaving 1,141 subjects for analysis. The ADOS diagnoses for these data were as follows: non-ASD = 170, ASD = 119, and autism = 919.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

audio (text transcript), we used the well-established method of automatic forced alignment of text to speech (Katsamanis, Black, Georgiou, Goldstein, & Narayanan, 2011).

The sessions were first manually transcribed through use of a protocol adapted from the Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2008) transcription guidelines and were segmented by speaker turn (i.e., the start and end times of each utterance in the acoustic waveform). The enriched transcription included partial words, stuttering, fillers, false starts, repetitions, nonverbal vocalizations, mispronunciations, and neologisms. Speech that was inaudible due to background noise was marked as such. In this study, speech segments that were unintelligible or that contained high background noise were excluded from further acoustic analysis.

With the lexical transcription completed, we then performed automatic phonetic forced alignment to the speech waveform using the HTK software (Young, 1993). Speech processing applications require that speech be represented by a series of acoustic features. Our alignment framework used the standard Mel-frequency cepstral coefficient (MFCC) feature vector, a popular signal representation derived from the speech spectrum, with standard HTK settings: 39-dimensional MFCC feature vector (energy of the signal + 12 MFCCs, and first- and second-order temporal derivatives), computed over a 25-ms window with a 10-ms shift. *Acoustic models* (AMs) are statistical representations of the sounds (phonemes) that make up words, based on the training data. Adult-speech AMs (for the psychologist's speech) were trained on the Wall Street Journal Corpus (Paul & Baker, 1992), and child-speech AMs (for the child's speech) were trained on the Colorado University (CU) Children's Audio Speech Corpus (Shobaki, Hosom, & Cole, 2000). The end result was an estimate of the start and end time of each phoneme (and, thus, each word) in the acoustic waveform.

Pitch and volume: Intonation and volume contours were represented by log-pitch and vocal intensity (short-time acoustic energy) signals that were extracted per word at turn-end using Praat software (Boersma, 2001). Pitch and volume contours were extracted only on turn-end words because intonation is most perceptually salient at phrase boundaries; in this work, we define the *turn-end* as the end of a speaker utterance (even if interrupted). In particular, turn-end intonation can indicate pragmatics such as disambiguating interrogatives from imperatives (Cruttenden, 1997), and it can indicate affect because pitch variability is associated with vocal arousal (Busso, Lee, & Narayanan, 2009; Juslin & Scherer, 2005). Turn-taking in interaction can lead to rather intricate prosodic display (Wells & MacFarlane, 1998). In this study, we examined multiple parameters of prosodic turn-end dynamics that may shed some light on the functioning of communicative intent. Future work could view complex aspects of prosodic functions through more precise analyses.

In this work, several decisions were made that may affect the resulting pitch contour statistics. Turns were included even if they contained overlapped speech, provided that the speech was intelligible. Thus, overlapped speech presented a potential source of measurement error. However, no significant relation was found between percentage overlap and ASD severity (p = 0.39), indicating that this may not have significantly affected results. Furthermore, we took an additional step to create more robust extraction of pitch. Separate

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

audio files were made that contained only speech from a single speaker (using transcribed turn boundaries); audio that was not from a target speaker's turns was replaced with Gaussian white noise. This was done in an effort to more accurately estimate pitch from the speaker of interest in accordance with Praat's pitch-extraction algorithm. Specifically, Praat uses a postprocessing algorithm that finds the cheapest path between pitch samples, which can affect pitch tracking when speaker transitions are short.

We investigated the dynamics of this turn-end intonation because the most interesting social functions of prosody are achieved by relative dynamics. Further, static functionals such as mean pitch and vocal intensity may be influenced by various factors unrelated to any disorder. In particular, mean pitch is affected by age, gender, and height, whereas mean vocal intensity is dependent on the recording environment and a participant's physical positioning. Thus, in order to factor variability across sessions and speakers, we normalized log-pitch and intensity by subtracting means per speaker and per session (see Equations 1 and 2). *Log-pitch* is simply the logarithm of the pitch value estimated by Praat; log-pitch (rather than linear pitch) was evaluated because pitch is log-normally distributed, and log-pitch is more perceptually relevant (Sonmez et al., 1997). Pitch was extracted with the autocorrelation method in Praat within the range of 75–600 Hz, using standard settings apart from minor empirically motivated adjustments (e.g., the octave jump cost was increased to prevent large frequency jumps):

$$\overline{f}_{0_{\log}}=\log(f_0)-E[\log(f_0)] \quad (1)$$

and

$$I\overline{n}t_v=Int_v-E[Int_v] \quad (2)$$

In order to quantify dynamic prosody, a second-order polynomial representation of turn-end pitch and vocal intensity was calculated that produced a curvature (2nd coefficient), slope (1st coefficient), and center (0th coefficient). *Curvature* measured rise–fall (negative) or fall–rise (positive) patterns; *slope* measured increasing (positive) or decreasing (negative) trends; and *center* roughly measured the signal level or mean. However, all three parameters were simultaneously optimized to reduce mean-squared error and, thus, were not exactly representative of their associated meaning. First, the time associated with an extracted feature contour was normalized to the range [−1,1] to adjust for word duration. An example parameterization is given in Figure 1 for the word *drives*. The pitch had a rise–fall pattern (curvature = −0.11), a general negative slope (slope = −0.12), and a positive level (center = 0.28).

Medians and interquartile ratios (IQRs) of the word-level polynomial coefficients representing pitch and vocal intensity contours were computed, totaling 12 features (2 Functionals × 3 Coefficients × 2 Contours). *Median* is a robust analogue of mean, and *IQR* is a robust measure of variability; functionals that are robust to outliers are advantageous, given the increased potential for outliers in this automatic computational study.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Rate: Speaking rate was characterized as the median and IQR of the word-level syllabic speaking rate in an utterance—done separately for the turn-end words—for a total of four features. Separating turn-end rate from non-turn-end rate enabled detection of potential affective or pragmatic cues exhibited at the end of an utterance (e.g., the psychologist could prolong the last word in an utterance as part of a strategy to engage the child). Alternatively, if the speaker were interrupted, the turn-end speaking rate might appear to increase, implicitly capturing the interlocutor's behavior.

Voice quality: Perceptual depictions of odd voice quality have been reported in studies of children with autism, having a general effect on the listenability of the children's speech. For example, children with ASD have been observed to have hoarse, harsh, and hypernasal voice quality and resonance (Pronovost, Wakstein, & Wakstein, 1966). However, interrater and intrarater reliability of voice quality assessment can vary greatly (Gelfer, 1988; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Thus, acoustic correlates of atypical voice quality may provide an objective measure that informs the child's ASD severity. Recently, Boucher et al. (2011) found that higher absolute jitter contributed to perceived "overall severity" of voice in spontaneous-speech samples of children with ASD. In this study, voice quality was captured by eight signal features: median and IQR of jitter, shimmer, cepstral peak prominence (CPP), and harmonics-to-noise ratio (HNR).

Jitter and *shimmer* measure short-term variation in pitch period duration and amplitude, respectively. Higher values for jitter and shimmer have been linked to perceptions of breathiness, hoarseness, and roughness (McAllister, Sundberg, & Hibi, 1998). Although speakers may hardly control jitter or shimmer voluntarily, it is possible that spontaneous changes in a speaker's internal state are indirectly responsible for such short-term perturbations of frequency and amplitude characteristics of the voice source activity. As reference, jitter and shimmer have been shown to capture vocal expression of emotion, having demonstrable relations with emotional intensity and type of feedback (Bachorowski & Owren, 1995) as well as stress (Li et al., 2007). In addition, whereas jitter and shimmer are typically only computed on sustained vowels when assessing dysphonia, jitter and shimmer are often informative of human behavior (e.g., emotion) in automatic computational studies of spontaneous speech; this is evidenced by the fact that jitter and shimmer are included in the popular speech processing tool kit openSMILE (Eyben, Wöllmer, & Schuller, 2010). In this study, modified variants of jitter and shimmer were computed that did not rely on explicit identification of cycle boundaries. Equation 3 shows the standard calculation for relative, local jitter, where T is the pitch period sequence and N is the number of pitch periods; the calculation of shimmer was similar and corresponded to computing the average absolute difference in vocal intensity of consecutive periods. In our study, smoothed, longer-term measures were computed by taking pitch period and amplitude samples every 20 ms (with a 40-ms window); the pitch period at each location was computed from the pitch estimated using the autocorrelation method in Praat. Relative, local jitter and shimmer were calculated on vowels that occurred anywhere in an utterance:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\begin{aligned} jitter_{loc,rel} &= \frac{jitter_{loc}}{mean\ Period}; \\ jitter_{loc} &= \frac{\sum_{j=2}^{N} |T_j - T_{j-1}|}{N-1}; \\ meanPeriod &= \sum_{j=1}^{N} \frac{T_j}{N}. \end{aligned} \tag{3}$$

CPP and HNR are measures of signal periodicity (whereas jitter is a measure of signal aperiodicity) that have also been linked to perceptions of breathiness (Hillenbrand, Cleveland, & Erickson, 1994) and harshness (Halberstam, 2004). For sustained vowels, percent jitter can be equally effective in measuring harshness as CPP in sustained vowels (Halberstam, 2004); however, CPP was even more informative when utilized on continuous speech. Heman-Ackah et al. (2003) found that CPP provided somewhat more robust measures of overall dysphonia than did jitter, when using a fixed-length windowing technique on read speech obtained at a 6-in. mouth-to-microphone distance. Because we worked with far-field (approximately 2-m mouth-to-microphone distance) audio recordings of spontaneous speech, voice quality measures may have been less reliable. Thus, we incorporated all four descriptors of voice quality, totaling eight features. We calculated HNR (for 0–1500 Hz) and CPP using an implementation available in VoiceSauce (Shue, Keating, Vicenik, & Yu, 2010); the original method was described in Hillenbrand et al. (1994) and Hillenbrand and Houde (1996). Average CPP was taken per vowel. Then, median and IQR (variability) of the vowel-level measures were computed per speaker as features (as done with jitter and shimmer).

Additional features: The style of interaction (e.g., who is the dominant speaker or the amount of overlap) may be indicative of the child’s behavior. Thus, we extracted four additional proportion features that represented disjoint segments of each interaction: (a) the fraction of the time in which the child spoke and the psychologist was silent, (b) the fraction of the time in which the psychologist spoke and the child was silent, (c) the fraction of the time that both participants spoke (i.e., “overlap”), and (d) the fraction of the time in which neither participant spoke (i.e., “silence”). These features were examined only in an initial statistical analysis.

Statistical Analysis

Spearman’s nonparametric correlation between continuous speech features and the discrete ADOS severity score was used to establish significance of relationships. Pearson’s correlation was used when comparing two continuous variables. The statistical significance level was set at $p < .05$. However, for the reader’s consideration, we sometimes report p values that did not meet this criterion but that, nonetheless, may represent trends that would be significant with a larger sample size (i.e., $p < .10$). In addition, underlying variables (e.g., psychologist identity, child age and gender, and signal-to-noise ratio [SNR; defined later in this paragraph]) were often controlled by using partial correlation in an effort to affirm significant correlations. *SNR* is a measure of the speech-signal quality affected by recording conditions (e.g., background noise, vocal intensity, or recorder gain). SNR was calculated as the relative energy within utterance boundaries (per speaker), compared with the energy in regions exclusive of utterance boundaries for either speaker.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Stepwise regression was performed on the entire data set in order to assess explanatory power through adjusted R^2 as well as examine selected features. Hierarchical and predictive regressions were performed to compare the explanatory power of the child's and the psychologist's acoustic-prosodic features. Given the limited sample size, stepwise feature selection was performed for all regressions. Parameters for stepwise regression were fixed for the stepwise regression and hierarchical regression sections (p_{intro} = .05 and p_{remove} = .10), and optimized for predictive regression.

Predictive regression was completed with a cross-validation framework to assess the model's explanatory power on an independent set of data; in particular, one session was held out for prediction, whereas the stepwise regression model was trained on all other sessions. The process was repeated in order to obtain a prediction for each session's severity rating. Then, the predicted severity ratings were correlated with the true severity ratings. All models included for selection the underlying variables (psychologist identity, age, gender, and SNR) in order to ensure that no advantage was given to either feature set. Parameters of stepwise regression were optimized per cross-fold; p_{intro} was selected in the range of [0.01, 0.19], with $p_{remove} = 2p_{intro}$.

Results

Relationship Between Normalized Speaking Times and Symptom Severity

Figure 2 illustrates the proportion of time spent talking by each participant, as well as periods of silence and overlapping speech. Correlations between duration of speech and ADOS severity are analyzed. The percentage of child speech (audible or inaudible due to background noise) during this subsample of the ADOS was not significantly correlated with ASD severity, $r_s(26) = -0.37$, $p = .06$. The percentage of psychologist speech was significantly correlated with ASD severity, $r_s(26) = 0.40$, $p = .03$. No relationship was found for percentage overlap ($p = .39$) or percentage silence ($p = .45$). Thus, the data suggest a pattern in which more frequent psychologist speech occurs with more severe ASD symptoms.

Child–Psychologist Coordination of Prosody

Certain prosodic features may co-vary between participants, suggesting that one speaker's vocal behavior is influenced by the other speaker's vocal behavior, or vice versa. The strongest correlation between participants was seen for median slope of vocal intensity, $r_p(26) = 0.64$, $p < .01$, as illustrated in Figure 3. This correlation was still significant at the $p < .01$ level after controlling for psychologist identity and SNR—presumably, the most likely confounding factors. Coordination of median jitter was not significant ($p = 0.24$), whereas coordination with median HNR was significant, $r_p(26) = .71$, $p < .001$, as displayed in Figure 4. Median jitter and HNR capture aspects of voice quality and can be altered unconsciously to some degree, although they are speaker dependent. After controlling for psychologist identity and SNR, significance at the $p = .05$ level was reached for median jitter, $r_p(26) = 0.47$, $p = .02$, as shown in Figure 5, and still existed for median HNR, $r_p(26) = 0.70$, $p < .001$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Two other features showed significant coordination between speakers: the pitch center IQRs and the CPP medians. But these relations were nonsignificant when controlling for psychologist identity and SNR, and thus were disregarded.

Relationship Between Acoustic-Prosodic Descriptors and ASD Severity

Correlation of acoustic-prosodic descriptors with ASD severity—In this subsection, the pairwise correlations between the 24 child and psychologist prosodic features and the rated ADOS severity are presented (see Table 1). Positive correlations indicate that increasing descriptor values corresponded to increasing symptom severity. If not stated otherwise, all reported correlations were still significant at the $p < .05$ significance level after controlling for the underlying variables: psychologist identity, age, gender, and SNR.

The pitch features of intonation were examined first. The child’s turn-end median pitch slope was negatively correlated with rated severity, $r_s(26) = −0.68$, $p < .001$; children with higher ADOS severity tended to have more negatively sloped pitch. Negative turn-end pitch slope is characteristic of statements, but also is related to other communicative functions such as turn-taking. Whether or not this acoustic feature may be associated with perceptions of monotonous speech is an area for further research. The child’s turn-end median pitch curvature showed similar correlations and could also be a marker of statements. In addition, the psychologist’s pitch center variability (IQR) was positively correlated with rated severity, $r_s(26) = 0.48$, $p < .01$, as was the psychologists’ pitch slope variability, $r_s(26) = 0.43$, $p < .05$; a psychologist tended to have more varied pitch center and pitch slope when interacting with a child who showed more atypical behavior. However, psychologist pitch center and slope variability correlations were nonsignificant ($p = .08$ and $p = .07$, respectively) after controlling for underlying variables; therefore, these results should be interpreted cautiously.

Next, we considered the vocal intensity features that describe intonation and volume. Psychologists’ vocal intensity center variability (IQR) was positively correlated with rated severity, $r_s(26) = 0.41$, $p = .03$. When interacting with a child whose behavior was more atypical, the psychologist tended to vary speech volume level more. Both the psychologist’s and the child’s vocal intensity slope variability (IQR) did not reach statistically significant positive correlation with ADOS severity ($p = .09$ and $p = .06$, respectively).

When examining speaking rate features, we observed qualitatively that some children with more severe symptoms spoke extremely fast, whereas others spoke extremely slow. The heterogeneity is consistent with the finding of no correlation between either speaker’s speaking rate features and the child’s rated severity.

Regarding measures of voice quality, we found several congruent relations with ADOS severity. Children’s median jitter was positively correlated with rated severity of ASD at $r_s(26) = 0.38$ ($p < .05$), whereas median HNR was negatively correlated at $r_s(26) = −0.38$ ($p < .05$); however, median CPP was not significantly correlated, $r_s(26) = −0.08$, $p = .67$. As a reminder, jitter is a measure of pitch aperiodicity, whereas HNR and CPP are measures of signal periodicity, and thus jitter is expected to have the opposite relations as HNR and CPP.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Similar to the child's features, the psychologist's median jitter, $r_S(26) = 0.43$, $p < .05$; median HNR, $r_S(26) = -0.37$, $p < .05$; and median CPP, $r_S(26) = -0.39$, $p < .05$, all indicate lower periodicity for increasing ASD severity of the child. Additionally, there were medium-to-large correlations for the child's jitter and HNR variability, $r_S(26) = 0.45$, $p < .05$, and $r_S(26) = 0.50$, $p < .01$, respectively, and for the psychologist's jitter, $r_S(26) = 0.48$, $p < .01$; CPP, $r_S(26) = 0.67$, $p < .001$; and HNR variability, $r_S(26) = 0.58$, $p < .01$—all indicate that increased periodicity variability is found when the child has higher rated severity. All of these voice quality feature correlations existed after controlling for the listed underlying variables, including SNR.

Stepwise regression—Stepwise multiple linear regression was performed using all child and psychologist acoustic-prosodic features as well as the underlying variables: psychologist identity, age, gender, and SNR to predict ADOS severity (see Table 2). The stepwise regression chose four features: three from the psychologist and one from the child. Three of these features were among those most correlated with ASD severity, indicating that the features contained orthogonal information. A child's negative pitch slope and a psychologist's CPP variability, vocal intensity center variability, and pitch center median all are indicative of a higher severity rating for the child according to the regression model. None of the underlying variables were chosen over the acoustic-prosodic features.

Hierarchical regression—In this subsection, we present the result of first optimizing a model for either the child's or the psychologist's features; then, we analyze whether orthogonal information is present in the other participant's features or the underlying variables (see Table 3); the included underlying variables are psychologist identity, age, gender, and SNR.

The same four features selected in the stepwise regression experiment were included in the child-first model, the only difference being that the child's pitch slope median was selected before the psychologist's CPP variability in this case. The child-first model only selected one child feature—child pitch slope median—and reached an adjusted R^2 of .43. Yet, further improvements in modeling were found ($R^2 = .74$) after selecting three additional psychologist features: (a) CPP variability, (b) vocal intensity center variability, and (c) pitch center median. A negative pitch slope for the child suggests flatter intonation, whereas the selected psychologist features may capture increased variability in voice quality and intonation.

The other hierarchical model first selects from psychologist features, then considers adding child and underlying features. That model, however, found that no significant explanatory power was available in the child or underlying features, with the psychologist's features contributing to an adjusted R^2 of .78. In particular, the model consists of four psychologist features: (a) CPP variability, (b) HNR variability, (c) jitter variability, and (d) vocal intensity center variability. These features largely suggest that increased variability in the psychologist's voice quality is indicative of higher ASD for the child.

Predictive regression—The results shown in Table 4 indicate the significant prediction of ADOS severity from acoustic-prosodic features. The psychologist's prosodic features

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

provided higher correlation than the child's prosodic features, $r_{s,psych}(26) = 0.79$, $p < .001$, compared with $r_{s,child}$ (26) = 0.64, $p < .001$, although the difference between correlations was not significant. Additionally, no improvement was observed when including the child's features for regression, $r_{s,psych\&child}$ (26) = 0.67, $p < .001$.

Discussion

The contributions of this work are threefold. First, semiautomatic processing and quantification of acoustic-prosodic features of the speech of children with ASD was conducted, demonstrating the feasibility of this paradigm for speech analysis even in the challenging domain of spontaneous dyadic interactions and the use of far-field sensors. Second, the unique approach of analyzing the psychologist's speech in addition to the child's speech during each interaction provided novel information about the predictive importance of the psychologist as an interlocutor in characterizing a child's autistic symptoms. Third, as predicted, speech characteristics of both the child and the psychologist were significantly related to the severity of the child's autism symptoms. Moreover, some proposed features such as intonation dynamics are novel to the ASD domain, whereas vocal quality measurements (e.g., jitter) mirrored other preliminary findings.

Examination of speaking duration indicated that the percentage of time in which the psychologist spoke in conversation was informative; in interactions with children who have more severe autism symptoms, the psychologist spoke more, and the child spoke nonsignificantly less ($p = .06$). This finding may suggest that the child with more severe ASD has difficulty conversing about the emotional and social content of the interview, and thus the psychologist is attempting different strategies, questions, or comments to try to draw the child out and elicit more verbal responses. Similar findings about relative speaking duration have been reported in previous observational studies of the interactions of adults and children or adolescents with autism (García-Perez, Lee, & Hobson, 2007; Jones & Schwartz, 2009). In addition, some coordination between acoustic-prosodic features of the child and the psychologist was shown for vocal intensity level variability, median HNR, and median jitter (only after controlling for underlying variables); this gives evidence of the interdependence of participants' behaviors. Vocal intensity is a significant contributor to perceived intonation, and HNR and jitter are related to aspects of atypical vocal quality. These findings suggest that, during the interactions, the psychologist tended to match her volume variability and voice quality to that of the child.

As predicted, correlation analyses demonstrated significant relationships between acoustic-prosodic features of both partners and rated severity of autism symptoms. Continuous behavioral descriptors that co-vary with this dimensional rating of social-affective behavior may lead to better phenotypic characterizations that address the heterogeneity of ASD symptomatology. Severity of autistic symptoms was correlated with children's negative turn-end pitch slope, which is a marker of statements. The underlying reason for this relationship is currently uncertain and needs further investigation. Children's jitter median tended to increase while HNR median decreased; jitter, HNR, and CPP variability also tended to increase in the children's speech with increasing ASD severity. Higher jitter, lower HNR, and lower CPP have been reported to occur with increased breathiness, hoarseness,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and roughness (Halberstam, 2004; Hillenbrand et al., 1994; McAllister et al., 1998), whereas similar perceptions of atypical voice quality have been reported in children with ASD. For example, Pronovost et al. (1966) found speakers with HFA to have hoarse, harsh, and hypernasal qualities. Hence, the less periodic values of jitter and HNR seen for children with higher autism severity scores suggest that the extracted measures are acoustic correlates of perceived atypical voice quality. The findings show promise for automatic methods of analysis, but there is uncertainty regarding which aspect of voice quality that jitter, HNR, and CPP may be capturing. Because the CPP measure was nonsignificant for the child, whereas the jitter and HNR measures were significant, further, more controlled investigation of voice quality during interaction is desired in future studies. The results corroborate findings from another acoustic study (Boucher et al., 2011), which found that higher absolute jitter contributed to perceived “overall severity” in samples of spontaneous speech of children with ASD.

Examination of the psychologist’s speech features revealed that when interacting with a more atypical child, the psychologist tended to vary her volume level and pitch dynamics (slope and center) more. This variability may reflect the psychologist’s attempts to engage the child by adding affect to her speech because increased pitch variability is associated with increased arousal (Juslin & Scherer, 2005). However, the pitch dynamic variability was nonsignificant (p = .08 and p = .07) after controlling for underlying variables, so this result should be interpreted with caution. It is also important to note that the data clearly show that certain relations are very significant and others should be further investigated with a more powerful clinical sample. Additionally, psychologist speech showed increased aperiodicity (captured by median jitter, CPP, and HNR) when interacting with children with higher autism severity ratings. This increased aperiodicity when interacting with more children who show more atypical behavior—together with the coordination observed between the two participants’ median HNR as well as their median jitter after controlling for underlying variables—suggests that the psychologist may be altering her voice quality to match that of the child. Furthermore, the psychologist’s periodicity variability (captured by jitter), CPP, and HNR variability—like the child’s—increased as the severity of autistic symptoms increased. Findings regarding voice quality are stronger for having considered several alternative measures.

Lastly, this study represents one of the first collections of empirical results that demonstrate the significance of psychologist behavior in relation to the severity of a child’s autism symptoms. In particular, three regression studies were conducted in this regard: stepwise regression, hierarchical stepwise regression, and predictive stepwise regression. *Stepwise regression* with selection from all child and psychologist acoustic-prosodic features and underlying variables demonstrated that both psychologist and child features had explanatory power for autism severity. *Hierarchical-stepwise regression* showed that, independently, both the child’s and the psychologist’s acoustic-prosodic features were informative. However, evidence suggests that the psychologist’s features were more explanatory than the child’s; higher R^2 was observed when selecting from the psychologist’s features than when selecting from the child’s features, and no child feature was selected after choosing from psychologist features first. Finally, the predictive value of each feature set was evaluated.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

The psychologist's features were more predictive of autism severity than were the child's features; although this difference was nonsignificant, the findings indicate that the psychologist's behavior carries valuable information about these dyadic interactions. Furthermore, the addition of the child's features to the psychologist's features did not improve prediction accuracy.

Implications for Future Research, Diagnosis, and Intervention

Two important results emerged in this study: First, the psychologist's prosody was at least as informative as the child's prosody of autism severity. Second, the semiautomatically extracted acoustic-prosodic features taken from spontaneous interactions between child and psychologist were correlated with autism severity. Future research could focus on sequential analysis of the psychologist's speech in order to gain more insights into the interaction dynamics between the child and the psychologist. For instance, it is of some interest to understand the point at which the psychologist makes a decision; this computation has been attempted in the couples therapy setting (Lee, Katsamanis, Georgiou, & Narayanan, 2012). Further, interaction processes such as prosodic entrainment can be computationally investigated in relation to expert-coded behaviors to lend deeper insights into underlying mechanisms (Lee et al., 2014). In addition, it would be helpful to sequentially analyze changes in the child's speech and level of engagement over the course of a session and whether these vary with changes in the psychologist's speech characteristics (Bone, Lee, & Narayanan, 2012).

Regarding the significance of the extracted acoustic-prosodic features, future research may investigate more specifically the relationship between prosody and overall ASD behavior impairments. Future research will also examine the prevalence of various prosodic abnormalities in children with a wider range of ASD severity and level of language functioning using computational techniques explored in this study but scaled to larger data sets. Dependencies of various prosodic abnormalities may also be examined, such as the effects of varying social and cognitive load throughout an interaction. Our recent preliminary work—which incorporates ratings of social load on the child—further investigates conversational quality by incorporating turn-taking and language features while expanding the analysis to the entire ADOS session (Bone et al., 2013). Greater understanding of the intricacies of atypical speech prosody can inform diagnosis and can lead to more personalized intervention. In addition, examination of children's specific responses to varied speech characteristics in the interacting partner may lead to fine-tuned recommendations for intervention targets and evaluation of mechanisms of change in intervention.

Conclusion

A framework was presented for objective, semiautomatic computation and quantification of prosody using speech signal features; such quantification may lead to robust prevalence estimates for various prosodic abnormalities and thus more specific phenotypic analyses in autism. Results indicate that the extracted speech features of both participants were informative. The extracted prosodic features were analyzed jointly with a dimensional rating

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

of social-affective behavior, motivated by the continuity of heterogeneous ASD symptomatology. Regression analyses provided empirical support for the significance of the psychologist's behavior in ASD assessment, an intuitive result given the dependence between dyadic interlocutors in general. These results support future study of acoustic prosody during spontaneous conversation—not only of the child's behavior but also of the psychologist's speech patterns—using computational methods that allow for analysis on much larger corpora. This preliminary study suggests that signal processing techniques have the potential to support researchers and clinicians with quantitative description of qualitative behavioral phenomena and to facilitate more precise stratification within this spectrum disorder.

Acknowledgments

This research was supported by funds from the National Science Foundation and by National Institutes of Health Grant PL-R21 HD065289. We are grateful to Sylvia Acosta and Irina Zamora for their administration of the ADOS and to Marcia Higareda for her recruiting efforts.

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4. Washington, DC: Author; 2000. text rev

American Speech-Language-Hearing Association. Childhood apraxia of speech (Position statement). 2007a. Retrieved from www.asha.org/policy/PS2007-00277.htm

American Speech-Language-Hearing Association. Childhood apraxia of speech (Technical report). 2007b. Retrieved from www.asha.org/policy/TR2007-00278.htm

Bachorowski JA, Owren MJ. Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. Psychological Science. 1995; 6:219–224.

Baltaxe C, Simmons JQ, Zee E. Intonation patterns in normal, autistic, and aphasic children. Proceedings of the Tenth International Congress of Phonetic Sciences. 1984:713–718.

Baron-Cohen S. Social and pragmatic deficits in autism: Cognitive or affective? Journal of Autism and Developmental Disorders. 1988; 18:379–402. [PubMed: 3049519]

Black MP, Bone D, Williams ME, Gorrindo P, Levitt P, Narayanan SS. The USC CARE Corpus: Child-psychologist interactions of children with autism spectrum disorders. Proceedings of Interspeech. 2011; 2011:1497–1500.

Boersma P. Praat: A system for doing phonetics by computer. Glot International. 2001; 5:341–345.

Bone D, Black MP, Lee CC, Williams ME, Levitt P, Lee S, Narayanan S. Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist. Proceedings of Interspeech. 2012; 2012:1043–1046.

Bone D, Lee CC, Chaspari T, Black MP, Williams ME, Lee S, Narayanan S. Acoustic-prosodic, turn-taking, and language cues in child–psychologist interactions for varying social demand. Proceedings of Interspeech. 2013; 2013:2400–2404.

Bone D, Lee CC, Narayanan S. A robust unsupervised arousal rating framework using prosody with cross-corpus evaluation. Proceedings of Interspeech. 2012; 2012:1175–1178.

Boucher, MJ.; Andrianopoulos, MV.; Velleman, SL.; Keller, LA.; Pecora, L. Assessing vocal characteristics of spontaneous speech in children with autism. Paper presented at the American Speech-Language-Hearing Association Convention; San Diego, CA. 2011 Nov.

Busso C, Lee S, Narayanan S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Transactions on Audio, Speech, and Language Processing. 2009; 17:582–596.

Cruttenden, A. Intonation. Cambridge, United Kingdom: Cambridge University Press; 1997.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Dawson G, Rogers S, Munson J, Smith M, Winter J, Greenson J, Varley J. Randomized, controlled trial of an intervention for toddlers with autism: The Early Start Denver Model. Pediatrics. 2010; 125:e17–e34. [PubMed: 19948568]

Diehl JJ, Watson D, Bennetto L, McDonough J, Gunlogson C. An acoustic analysis of prosody in high-functioning autism. Applied Psycholinguistics. 2009; 30:385–404.

Eyben, F.; Wöllmer, M.; Schuller, B. OpenSMILE: The Munich versatile and fast open-source audio feature extractor. Proceedings of the 18th International Conference on Multimedia; ACM Multimedia; 2010. p. 1459-1462.

Frith U. Mind blindness and the brain in autism. Neuron. 2001; 32:969–980. [PubMed: 11754830]

Frith U, Happé F. Autism spectrum disorder. Current Biology. 2005; 15:R786–R790. [PubMed: 16213805]

Furrow D. Young children's use of prosody. Journal of Child Language. 1984; 11:203–213. [PubMed: 6699111]

García-Perez RM, Lee A, Hobson RP. On intersubjective engagement in autism: A controlled study of nonverbal aspects of communication. Journal of Autism and Developmental Disorders. 2007; 37:1310–1322. [PubMed: 17086439]

Gelfer MP. Perceptual attributes of voice: Development and use of rating scales. Journal of Voice. 1988; 2:320–326.

Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P, Spence SJ. The Autism Genetic Resource Exchange: A resource for the study of autism and related neuropsychiatric conditions. American Journal of Human Genetics. 2001; 69:463. [PubMed: 11452364]

Gotham K, Pickles A, Lord C. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. Journal of Autism and Developmental Disorders. 2009; 39:693–705. [PubMed: 19082876]

Gotham K, Risi S, Pickles A, Lord C. The Autism Diagnostic Observation Schedule: Revised algorithms for improved diagnostic validity. Journal of Autism and Developmental Disorders. 2007; 37:613–627. [PubMed: 17180459]

Halberstam B. Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. ORL. 2004; 66:70–73. [PubMed: 15162004]

Heman-Ackah YD, Heuer RJ, Michael DD, Ostrowski R, Horman M, Barody MM, Sataloff RT. Cepstral peak prominence: A more reliable measure of dysphonia. Annals of Otology, Rhinology & Laryngology. 2003; 112:324–333.

Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. Journal of Speech, Language, and Hearing Research. 1994; 37:769–778.

Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. Journal of Speech, Language, and Hearing Research. 1996; 39:311–321.

Jones CD, Schwartz IS. When asking questions is not enough: An observational study of social communication differences in high functioning children with autism. Journal of Autism and Developmental Disorders. 2009; 39:432–443. [PubMed: 18784993]

Juslin, PN.; Scherer, KR. Vocal expression of affect. In: Harrigan, J.; Rosenthal, R.; Scherer, K., editors. The new handbook of methods in nonverbal behavior research. Oxford, United Kingdom: Oxford University Press; 2005. p. 65-135.

Katsamanis, A.; Black, MP.; Georgiou, PG.; Goldstein, L.; Narayanan, S. SailAlign: Robust long speech-text alignment. Paper presented at the Workshop on New Tools and Methods for Very-Large-Scale Phonetics Research; Philadelphia, PA: University of Pennsylvania; 2011 Jan.

Kimura M, Daibo I. Interactional synchrony in conversations about emotional episodes: A measurement by "the between-participants pseudosynchrony experimental paradigm. Journal of Nonverbal Behavior. 2006; 30:115–126.

Knapp, ML.; Hall, JA. Nonverbal communication in human interaction. Belmont, CA: Wadsworth; 2009.

Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. Journal of Speech, Language, and Hearing Research. 1993; 36:21–40.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Lee CC, Katsamanis A, Black MP, Baucom BR, Christensen A, Georgiou PG, Narayanan SS. Computing vocal entrainment: A signal-derived PCA-based quantification scheme with applications to affect analysis in married couple interactions. Computer Speech & Language. 2014; 28:518–539.

Lee CC, Katsamanis A, Georgiou PG, Narayanan SS. Based on isolated saliency or causal integration? Toward a better understanding of human annotation process using multiple instance learning and sequential probability ratio. Proceedings of Interspeech. 2012; 2012:619–622.

Li X, Tao J, Johnson MT, Soltis J, Savage A, Leong KM, Newman JD. Stress and emotion classification using jitter and shimmer features. Proceedings of IEEE ICASSP. 2007 Apr.:1081–1084.

Lord C, Jones RM. Annual research review: Rethinking the classification of autism spectrum disorders. Journal of Child Psychology and Psychiatry. 2012; 53:490–509. [PubMed: 22486486]

Lord, C.; Rutter, M.; DiLavore, PC.; Risi, S. Autism Diagnostic Observation Schedule. Los Angeles, CA: Western Psychological Services; 1999.

Lord, C.; Rutter, M.; DiLavore, PC.; Risi, S.; Gotham, K.; Bishop, SL. Autism Diagnostic Observation Schedule. 2. Torrance, CA: Western Psychological Services; 2012.

McAllister A, Sundberg J, Hibi SR. Acoustic measurements and perceptual evaluation of hoarseness in children’s voices. Logopedics Phonatrics Vocology. 1998; 23:27.

McCann J, Peppe S. Prosody in autism spectrum disorders: A critical review. International Journal of Language & Communication Disorders. 2003; 38:325–350. [PubMed: 14578051]

Miller, JF.; Iglesias, A. Systematic Analysis of Language Transcripts (English and Spanish, Version 9) [Computer software]. Madison, WI: University of Wisconsin—Madison, Waisman Center, Language Analysis Laboratory; 2008.

Narayanan S, Georgiou PG. Behavioral signal processing: Deriving human behavioral informatics from speech and language. Proceedings of IEEE. 2013; 101:1203–1233.

Paccia JM, Curcio F. Language processing and forms of immediate echolalia in autistic children. Journal of Speech, Language, and Hearing Research. 1982; 25:42–47.

Paul, DB.; Baker, JM. The design for the Wall Street Journal–based CSR corpus. Proceedings of the ACL Workshop on Speech and Natural Language; Stroudsburg, PA: Association for Computational Linguistics; 1992. p. 357-362.

Paul R, Augustyn A, Klin A, Volkmar FR. Perception and production of prosody by speakers with autism spectrum disorders. Journal of Autism and Developmental Disorders. 2005; 35:205–220. [PubMed: 15909407]

Paul R, Shriberg LD, McSweeny J, Cicchetti D, Klin A, Volkmar F. 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. 2005; 35:861–869. [PubMed: 16283080]

Peppe, S. Assessment of prosodic ability in atypical populations, with special reference to high-functioning autism. In: Stojanovik, V.; Setter, J., editors. Speech prosody in atypical populations: Assessment and remediation. Guildford, United Kingtom: J&R Press; 2011. p. 1-23.

Peppe S, McCann J, Gibbon F, O’Hare A, Rutherford M. Receptive and expressive prosodic ability in children with high-functioning autism. Journal of Speech, Language, and Hearing Research. 2007; 50:1015–1028.

Ploog BO, Banerjee S, Brooks PJ. Attention to prosody (intonation) and context in children with autism and in typical children using spoken sentences in a computer game. Research in Autism Spectrum Disorders. 2009; 3:743–758.

Prizant BM, Wetherby AM, Rubin MS, Laurent AC. The SCERTS model: A transactional, family-centered approach to enhancing communication and socioemotional abilities of children with autism spectrum disorder. Infants and Young Children. 2003; 16:296–316.

Pronovost W, Wakstein MP, Wakstein DJ. A longitudinal study of the speech behavior and language comprehension of fourteen children diagnosed atypical or autistic. Exceptional Children. 1966; 33:19–26. [PubMed: 5981255]

Rutter, M.; LeCouteur, A.; Lord, C. Autism Diagnostic Interview—Revised manual. Los Angeles, CA: Western Psychological Services; 2003.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Sheinkopf SJ, Mundy P, Oller DK, Steffens M. Vocal atypicalities of preverbal autistic children. Journal of Autism and Developmental Disorders. 2000; 30:345–354. [PubMed: 11039860]

Shobaki K, Hosom JP, Cole R. The OGI Kids' speech corpus and recognizers. Proceedings of ICLSP 2000. 2000; 4:258–261.

Shriberg LD, Austin D, Lewis BA, McSweeny JL, Wilson DL. The Speech Disorders Classification System (SDCS): Extensions and lifespan reference data. Journal of Speech, Language, and Hearing Research. 1997; 40:723–740.

Shriberg LD, Fourakis M, Hall SD, Karlsson HB, Lohmeier HL, McSweeny JL, Wilson DL. Extensions to the Speech Disorders Classification System (SDCS). Clinical Linguistics & Phonetics. 2010; 24:795–824. [PubMed: 20831378]

Shriberg LD, Paul R, Black LM, van Santen JP. The hypothesis of apraxia of speech in children with autism spectrum disorder. Journal of Autism and Developmental Disorders. 2011; 41:405–426. [PubMed: 20972615]

Shriberg LD, Paul R, McSweeny JL, Klin A, Cohen DJ, Volkmar FR. Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. Journal of Speech, Language, and Hearing Research. 2001; 44:1097–1115.

Shue YL, Keating P, Vicenik C, Yu K. VoiceSauce: A program for voice analysis. Energy. 2010; 1(H2):H1–A1.

Siller M, Sigman M. The behaviors of parents of children with autism predict the subsequent development of their children's communication. Journal of Autism and Developmental Disorders. 2002; 32:77–89. [PubMed: 12058846]

Sonmez, MK.; Heck, L.; Weintraub, M.; Shriberg, E.; Kemal, M.; Larry, S.; Shriberg, WE. A lognormal tied mixture model of pitch for prosody-based speaker recognition. Paper presented at the Fifth European Conference on Speech Communication and Technology; Rhodes, Greece. 1997 Sep.

Uldall E. Attitudinal meanings conveyed by intonation contours. Language and Speech. 1960; 3:223–234.

van Santen JP, Prud'hommeaux ET, Black LM, Mitchell M. Computational prosodic markers for autism. Autism. 2010; 14:215–236. [PubMed: 20591942]

Vernon TW, Koegel RL, Dauterman H, Stolen K. An early social engagement intervention for young children with autism and their parents. Journal of Autism and Developmental Disorders. 2012; 42:2702–2717. [PubMed: 22527708]

Weider S, Greenspan SI. Climbing the symbolic ladder in the DIR model through floor time/interactive play. Autism. 2003; 7:425–435. [PubMed: 14678681]

Wells B, MacFarlane S. Prosody as an interactional resource: Turn-projection and overlap. Language and Speech. 1998; 41:265–294. [PubMed: 10746359]

Young, SJ. Technical Report No 153. Cambridge, United Kingdom: Department of Engineering, Cambridge University; 1993. The HTK Hidden Markov Model toolkit: Design and philosophy.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

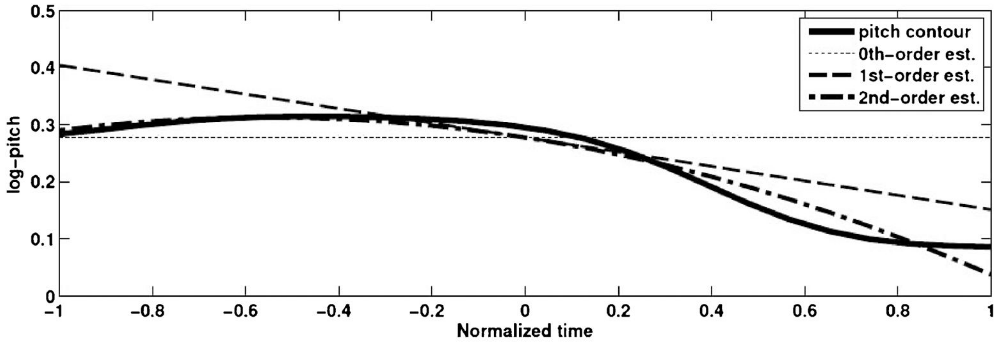

Figure 1.

Second-order polynomial representation of the normalized log-pitch contour for the word *drives.* Curvature = −0.11; slope = −0.12; center = 0.28.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

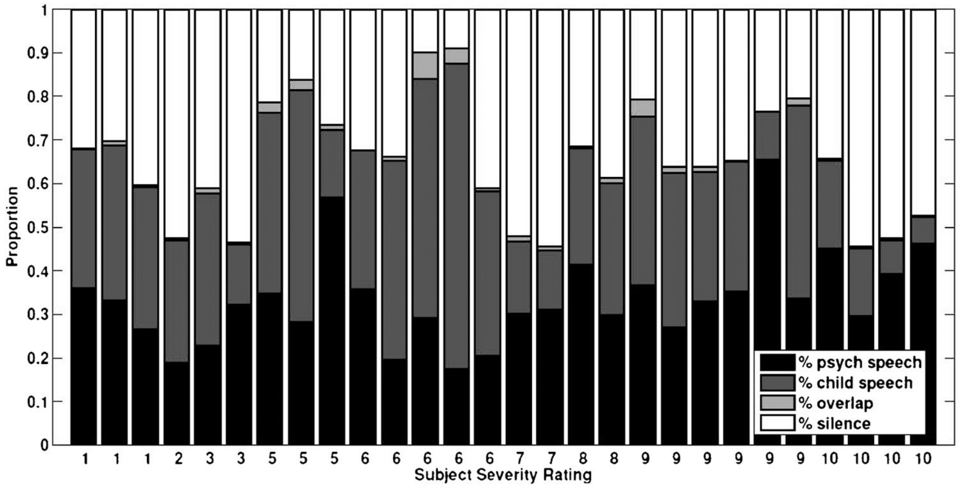

Figure 2.
Proportions of conversation containing psychologist and/or child speech. Sessions are ordered and labeled by Autism Diagnostic Observation Scale (ADOS) severity.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

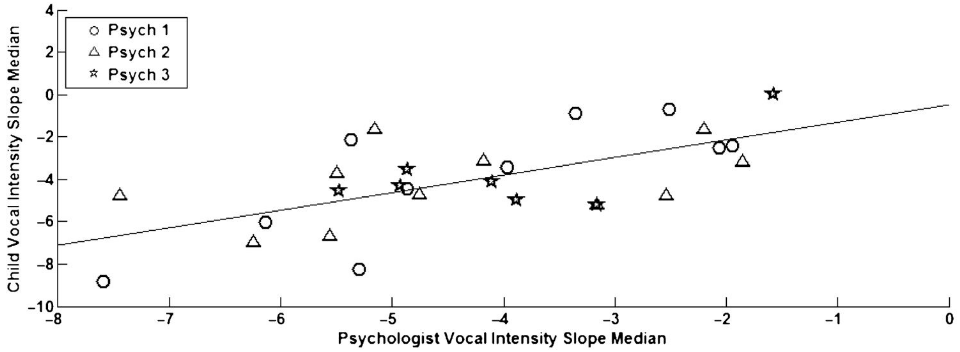

Figure 3.

Coordination of vocal intensity slope median between child and psychologist.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

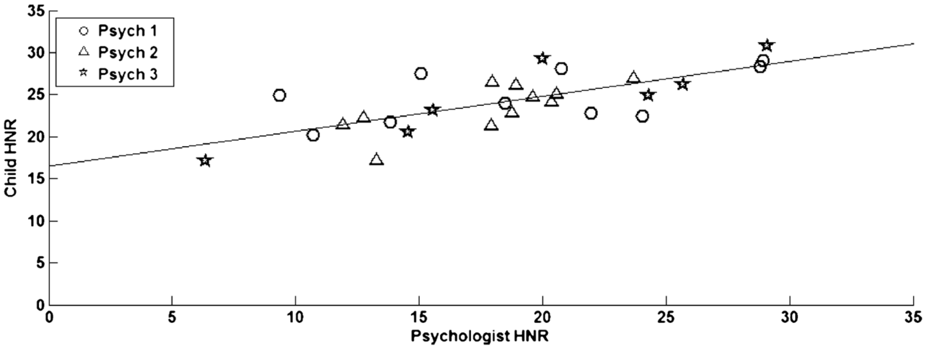

Figure 4.

Coordination of median harmonics-to-noise ratio (HNR) between child and psychologist.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

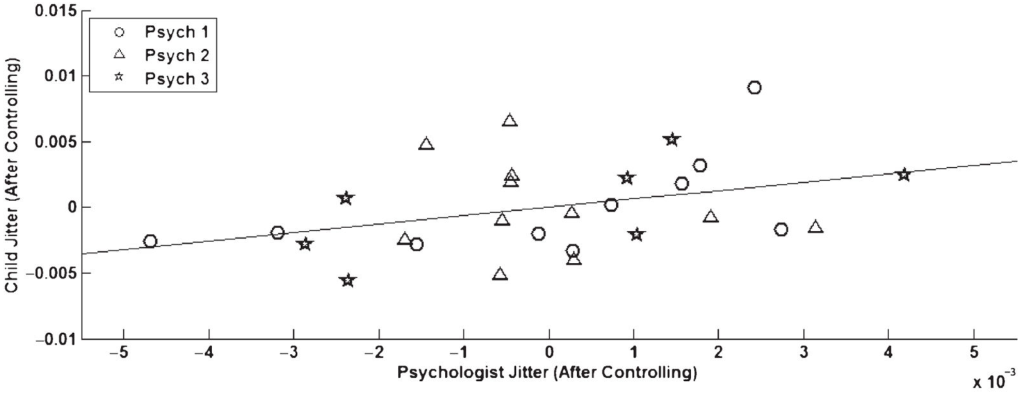

Figure 5.

Coordination of median jitter between child and psychologist after controlling for psychologist identity and signal-to-noise ratio (SNR).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Spearman rank order correlation coefficients between acoustic-prosodic descriptors and ADOS severity.

Category	Descriptor	Child r_s	Psychologist r_s
Intonation: Pitch	Curvature median	-0.53^{**}	−0.12
	Slope median	-0.68^{**}	0.30
	Center median	−0.12	0.26
	Curvature IQR	0.22	0.09
	Slope IQR	−0.03	0.43^{*}
	Center IQR	0.02	0.48^{**}
Intonation: Vocal intensity	Curvature median	−0.09	−0.13
	Slope median	−0.31	−0.25
	Center median	−0.14	0.09
	Curvature IQR	−0.05	0.10
	Slope IQR	$0.36^{\dagger}$	$0.33^{\dagger}$
	Center IQR	0.18	0.41^{*}
Speaking rate	Nonboundary median	−0.00	0.19
	Boundary median	0.00	−0.04
	Nonboundary IQR	0.22	−0.05
	Boundary IQR	$0.33^{\dagger}$	−0.03
Voice quality	Jitter median	0.38^{*}	0.43^{*}
	Shimmer median	0.08	0.04
	CPP median	−0.03	-0.39^{*}
	HNR median	-0.38^{*}	-0.37^{*}
	Jitter IQR	0.45^{*}	0.48^{**}
	Shimmer IQR	−0.12	−0.03
	CPP IQR	0.12	0.67^{***}
	HNR IQR	0.50^{**}	0.58^{**}

Note. Positive correlations indicate that increasing descriptor values occur with increasing severity. IQR = interquartile ratio; HNR = harmonics-to-noise ratio; CPP = cepstral peak prominence.

†*p* < .10.

**p* < .05.

***p* < .01.

****p* < .001.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Stepwise regression with prosodic features and underlying variables.

		Cumulative model statistics					Final βs	
Step	**Added feature**	***R***	R^2	**Adj.** R^2	***F***	***p***	**Standardized β**	***p***
1	Psychologist CPP IQR	0.71	0.50	.48	26.3	< .001	.50	< .001
2	Child pitch slope median	0.81	0.65	.62	23.3	< .001	−.33	< .01
3	Psychologist vocal int. center IQR	0.85	0.72	.68	20.3	< .001	.34	< .01
4	Psychologist pitch center median	0.88	0.78	.74	20.4	< .001	.26	.02

Note. Int. = intensity.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 3

Hierarchical stepwise regression with prosodic features of either participant and underlying variables.

		Cumulative model statistics					Final βs	
Step	**Added feature**	***R***	***R^2***	**Adj. *R^2***	***F***	***p***	**Standardized β**	***p***
		Child prosody, then psych prosody and underlying variables						
1	Child pitch slope median	.67	.45	.43	21.6	< .001	−0.33	< .01
2	Psych CPP IQR	.81	.65	.62	23.3	< .001	0.50	< .001
3	Psych vocal int. center IQR	.85	.72	.68	20.3	< .001	0.34	< .01
4	Psych pitch slope median	.88	.78	.74	20.4	< .001	0.27	.02
		Psych prosody, then child prosody and underlying variables						
1	Psych CPP IQR	.71	.50	.48	26.3	< .001	0.48	< .001
2	Psych HNR IQR	.79	.63	.60	21.2	< .001	0.32	< .01
3	Psych jitter IQR	.84	.71	.69	19.1	< .001	0.35	< .01
4	Psych vocal int. center IQR	.90	.81	.78	24.3	< .001	0.33	< .01
1	Psych CPP IQR	.71	.50	.48	26.3	< .001	0.48	< .001

Note. Variables are included for selection in this order: child's (or psychologist's) prosody, psychologist's (or child's) prosody, and underlying variables.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 4

Spearman rank order correlation between predicted severity based on acoustic-prosodic descriptors and actual, rated ADOS severity.

Descriptors included	Child prosody	Psych prosody	Child and psych prosody	Underlying variables
r_s	0.64^{***}	0.79^{***}	0.67^{***}	−0.14

*** $p < .001$.