Published in final edited form as:

Autism. 2010 May ; 14(3): 215–236. doi:10.1177/1362361309363281.

Computational Prosodic Markers for Autism

J. P. H. van Santen,
Oregon Health & Science University, USA

E. T. Prud'hommeaux,
Oregon Health & Science University, USA

L. M. Black, and
Oregon Health & Science University, USA

M. Mitchell
University of Aberdeen, Scotland, UK

Abstract

We present results obtained with new instrumental methods for the acoustic analysis of prosody to evaluate prosody production by children with Autism Spectrum Disorder (ASD) and Typical Development (TD). Two tasks elicit focal stress, one in a vocal imitation paradigm, the other in a picture-description paradigm; a third task also uses a vocal imitation paradigm, and requires repeating stress patterns of two-syllable nonsense words. The instrumental methods differentiated significantly between the ASD and TD groups in all but the focal stress imitation task. The methods also showed smaller differences in the two vocal imitation tasks than in the picture-description task, as was predicted. In fact, in the nonsense word stress repetition task, the instrumental methods showed better performance for the ASD group. The methods also revealed that the acoustic features that predict auditory-perceptual judgment are not the same as those that differentiate between groups. Specifically, a key difference between the groups appears to be a difference in the balance between the various prosodic cues, such as pitch, amplitude, and duration, and not necessarily a difference in the strength or clarity with which prosodic contrasts are expressed

Verbal communication has two aspects: What is said and how it is said. The latter refers to *prosody*, defined as the use of acoustic features of speech to complement, highlight, or modify the meaning of what is said. Among the best known of these features are fundamental frequency (F_0 , or, informally, pitch), duration (e.g., Klatt, 1976), and intensity (e.g., Fry, 1955); less known is spectral balance, which is a correlate of, for example, oral aperture and breathiness (e.g., Sluijter, Shattuck-Hufnagel, Stevens, & van Heuven, 1995; Campbell & Beckman, 1997; Sluijter, van Heuven, & Pacilly, 1997; van Santen & Niu, 2003; Miao, Niu, Klabbers, & van Santen, 2006). Exactly which acoustic features are involved in prosody and how they interact is an intrinsically complex issue that still is only partially understood. First, prosodic features tend to be used jointly. For example, the end of a phrase is typically signaled in the final one or two syllables by slowing down, lowering pitch, decreasing intensity, and increasing breathiness. Second, speakers may compensate for making less use of one feature by making more use of another feature (“cue trading”; Beach, 1991). Thus, at the end of a phrase some speakers may slow down more than others, but increase breathiness less. Third, while in certain cases (e.g., affective prosody) global features such as average loudness or pitch range are useful descriptors, in other cases (e.g.,

contrastive stress) subtle details of the dynamic patterns of these features, such as pitch peak timing (e.g., Post, d'Imperio, & Gussenhoven, 2007), are important.

Prosody plays a crucial role in an individual's communicative competence and social-emotional reciprocity. Expressive prosody deficits have been considered among the core features of autism spectrum disorders (ASD) in individuals who are verbal since Kanner first described the disorder (Kanner, 1943). The ability to appropriately understand and express prosody may, in fact, be an integral part of the theory of mind deficits considered central to autism, and may play a role in the reported lack of ability in individuals with ASD to make inferences about others' intentions, desires, and feelings (Sabbagh, 1999; Tager-Flusberg, 2000; Baron-Cohen, 2000).

Nevertheless, expressive prosody has not been one of the core diagnostic criteria for the disorder in most versions of the Diagnostic and Statistical Manual of Mental Disorders that included it (DSM-III [American Psychiatric Association, 1980] through DSM-IV-TR [American Psychiatric Association, 2002], with one exception: DSM-III-R [American Psychiatric Association, 1987]); nor has it been described as part of the broad neurobehavioral phenotype (Dawson, Webb, Schellenberg, Dager, Friedman, Aylward, et al., 2001); nor does it appear on any "algorithm" of the ADOS (Lord, Risi, Lambrecht, Cook, Leventhal, DiLavore, et al., 2000; Gotham, Risi, Pickles, & Lord, 2007), the standard instrument used in research for diagnosis of ASD. This absence of expressive prosody in diagnostic criteria and instruments is perhaps due to difficulties in its reliable measurement as well as to uncertainty about the numbers affected in the broad, heterogeneous spectrum. Recent research has, however, begun to document impairments. Prosody characteristics noted for persons with ASD have included monotonous speech as well as sing-song speech, aberrant stress, atypical pitch patterns, abnormalities of rate and volume of speech, and problematic quality of voice (e.g., Baltaxe, 1981; Baltaxe, Simmons, & Zee, 1984; Shriberg, Paul, McSweeney, Klin, Cohen, & Volkmar, 2001). Findings obtained with the Profiling Elements of Prosodic Systems-Children instrument (PEPS-C; Peppé & McCann, 2003), a computerized battery of tests consisting of decontextualized tasks that span the linguistic-pragmatic-affective continuum, include poorer performance by children with ASD compared to typically developing (TD) children, particularly on affective and pragmatic prosody tasks (Peppé, McCann, Gibbon, O'Hare, & Rutherford, 2007; McCann, Peppé, Gibbon, O'Hare, & Rutherford, 2007).

In combination, these findings indeed confirm the presence of expressive prosody impairments in ASD. These findings, however, are all based on auditory-perceptual methods and not on instrumental acoustic measurement methods. Auditory-perceptual methods have fundamental problems. First, the reliability and validity of auditory-perceptual methods is often lower than desirable (e.g., Kent, 1996; Kreiman & Gerratt, 1998), due to a variety of factors. For example, it is difficult to judge one aspect of speech without interference from other aspects (e.g., judging nasality in the presence of varying degrees of hoarseness); certain judgment categories are intrinsically multidimensional, thus requiring each judge to weigh subjectively and individually these dimensions; and there is a paucity of reference standards. Second, auditory-perceptual methods require human judges, which limits the amount of speech that can be practically processed, especially when reliability concerns necessitate a panel of judges. Third, tasks for assessing prosody typically are administered face-to-face, which makes it likely that the examiner's auditory-perceptual judgments are influenced by the diagnostic impressions that unavoidably result from observing the child's overall verbal and nonverbal behavior while performing the task. This is in particular the case when judging how atypical prosody may be in ASD, given the disorder's many other behavioral cues. Fourth, and most important, speech features detected by auditory-perceptual methods are by definition features that are audible, not overshadowed by other

features, and can be reliably (and separately from other features) coded by humans. However, there is no a priori reason why these particular features would be precisely those that could critically distinguish between diagnostic groups and serve as potential markers of specific disorders.

These considerations make it imperative to search for instrumental prosody assessment methods that are (i) objective, and hence, within the limits of how representative a speech sample is of an individual's overall speech, reliable; (ii) automated; (iii) capture a range of speech features, thereby enabling the discovery of markers that might be difficult for judges to perceive or code; and (iv) target these features separately and independently of one another.

In this paper, we describe and apply technology-based instrumental methods to explore differences in expressive prosody between children with ASD and children with typical development. These methods represent a new generation of instrumental methods that target dynamic patterns of individual acoustic features and are fully automatic. In contrast, current acoustic methods that have been applied to ASD (e.g., Baltaxe, 1984; Fosnot & Jun, 1999; Diehl, Watson, Bennetto, McDonough, & Gunlogson, 2009) capture only global features such as average pitch or pitch range instead of dynamic patterns; they also require manual labor.

These instrumental prosody assessment methods were introduced by van Santen and his co-workers (van Santen, Prud'hommeaux, & Black, 2009) and validated against an auditory-perceptual method. In this auditory-perceptual method, six listeners independently judged the direction and the strength of contrast in "prosodic minimal pairs" of recordings. Each prosodic minimal pair contained two utterances from the same child containing the same phonemic material but differing on a specific prosodic contrast, such as stress (e.g., '*tauveeb*' and '*tau'veeb*'). The averages of the ratings of the six listeners will be referred to as the "Listener scores." The instrumental methods use pattern-detection algorithms to compare the two utterances of a minimal pair in terms of their prosodic features. Van Santen et al. (2009) showed that these measures correlated with the Listener scores approximately as well as the judges' individual scores did, and substantially better than "Examiner scores" (i.e., scores assigned by examiners during assessment that could later be verified off-line). Examiner scores represent what can be reasonably expected in terms of effort and quality in clinical practice, while the process of obtaining the Listener scores is impractical and only serves research purposes. The fact that the instrumental measures correlated well with these high-quality scores has, of course, immediate practical significance. The instrumental methods will be described in detail in the Methods section.

The main goal of the current paper is to use these instrumental methods to explore differences between children with ASD and children with TD, in particular to test the following hypotheses that are suggested by the above discussion:

1. The proposed instrumental methods will differentiate between ASD and TD groups.
2. Performance on prosodic speech production tasks will be poorer in children with ASD compared to children with TD, whether measured with auditory-perceptual methods (i.e., Examiner scores) or with the instrumental methods. This performance difference will be particularly pronounced in tasks that require more than an immediate repetition of a spoken word, phrase, or sentence, such as tasks that entail describing pictorial materials and those involving pragmatic or affective prosody.

3. Prosodic features that, for a given task, are predictive of auditory-perceptual scores are not necessarily the same as those that differentiate between the groups.

We apply these methods in a research protocol that addresses several methodological shortcomings discussed by McCann and Peppé (2003) in their review of prosody in ASD, by using samples of individuals with ASD that (i) are well-characterized; (ii) matched on key variables such as non-verbal IQ and age; (iii) have adequate sizes (most studies reviewed by McCann & Peppé have samples of fewer than 10 individuals, and only three out of thirteen studies reviewed have samples larger than 15); and (iv) have an adequately narrow age range (e.g., the studies with the largest numbers of participants, regrettably, have wide age ranges: 7–32 years in Fine, J., Bartolucci, G., Ginsberg, G., & Szatmari (1991); 10–50 years in Shriberg et al. (2001); and, in a more recent study not covered by McCann & Peppé, 7–28 years in Paul, Bianchi, Augustyn, Klin, & Volkmar (2008)).

Participants

Subjects were children between the ages of four and eight years. The ASD group was “high functioning” (HFA), with a full scale IQ of at least 70 (Gillberg & Ehlers, 1998; Siegel, Minshew & Goldstein, 1996) and “verbally fluent”. Concerning the latter, a Mean Length of Utterance (MLU, measured in morphemes per utterance) of at least 3 was required to enter the study; in addition, during a clinical screening procedure, a Speech Language Pathologist verified whether a child’s expressive speech was adequate for performing the speech tasks in the protocol, and also judged intelligibility to ascertain that it was adequate for transcribing the speech recordings. The groups as recruited were initially not matched on nonverbal IQ (NVIQ) and age, with the TD group being younger and having a higher NVIQ. NVIQ was measured using the Wechsler Scales across the age range: for children 4.0–6.11 years, Performance IQ (PIQ) from the WPPSI-III (Wechsler Preschool and Primary Scale of Intelligence, third edition) was used; for children 7.0–8.11 years, the Perceptual Reasoning Index (PRI) from the WISC-IV (The Wechsler Intelligence Scale for Children, fourth edition) was used. Matching was achieved via a post-selection process that was unbiased with respect to diagnosis and based on no information other than age and NVIQ, and in which children were eliminated alternating between groups until the between-group difference in age and NVIQ was not significant at $p < 0.10$, one-tailed. In addition, in all analyses reported below, children were excluded whose scores on the acoustic features analyzed were outliers, defined as deviating by more than 2.33 standard deviations (i.e., the 1st or 99th percentile) from their group mean. These extreme deviations in acoustic features were generally due to task administration failures. Results reported below on the children thus selected did not substantially change when the matching or outlier criteria were made slightly less or more stringent; several results did change substantially and unpredictably, however, when matching was dropped, implying age- and/or NVIQ-dependency of the results. Future studies will analyze these dependencies in detail, as more data become available. In all analyses, the numbers of children ranged from 23 to 26 in the TD group and 24 to 26 in the ASD group.

Table I contains averages and standard deviations for age, NVIQ, and the three subtests (Block Design, Matrix Reasoning, and Picture Concepts) on which NVIQ is based. As can be seen, the samples are well matched for age and NVIQ. The table also shows that the TD group scored significantly higher than the ASD group on the Picture Concepts subtest, nearly the same on the Matrix Reasoning subtest, and somewhat lower on the Block Design subtest. This pattern may reflect the fact that of these three subtests the Picture Concepts subtest makes the heaviest verbal demands.

Groups were matched on NVIQ rather than on VIQ for two reasons. First, language difficulties are frequently present in ASD (e.g., Tager-Flusberg, 2000; Leyfer, Tager-

Flusberg, Dowd, Tomblin, & Folstein, 2008). Second, it is fairly typical in ASD (although not uniformly) that $NVIQ > VIQ$ (e.g., Joseph, Tager-Flusberg, & Lord, 2002). Our sample was no exception. Thus, matching on VIQ would have created an ASD group with a substantially higher NVIQ than the TD group; results from such a comparison would have been difficult to interpret. We also note that our key results are based on intra-individual comparisons between tasks, thereby creating built-in controls (Jarrold & Brock, 2004) that make results less susceptible to matching issues.

All children in the ASD group had received a prior clinical, medical, or educational diagnosis of ASD as a requirement for participation in the study. To confirm, however, whether a child indeed met diagnostic classification criteria for the ASD group, each child was further evaluated. The decision to classify a child in the ASD group was based on: (i) results on the revised algorithm of the ADOS (Lord et al., 2000; Gotham, Risi, Dawson, Tager-Flusberg, Joseph, Carter, et al., 2008), administered and scored by trained, certified clinicians; (ii) results on The Social Communication Questionnaire (SCQ; Berument, Rutter, Lord, Pickles, & Bailey, 1999), a parent-report measure; and (iii) results of a consensus clinical diagnosis made by a multidisciplinary team in accord with DSM-IV criteria (American Psychiatric Association, 2002). In all cases, a diagnosis of ASD was supported by ADOS scores at or above cutoffs for PDD-NOS and by the consensus clinical diagnosis. There were four cases in which the SCQ score was below the conventional cutoff of 12, but in each of these, the SCQ score was at least 9 (2 at 11, 2 at 9), and the diagnosis was strongly supported by consensus clinical judgment and ADOS algorithm scores. The percentage of cases that had been pre-diagnosed with ASD but did not meet the study's ASD inclusion criteria was about 15%.

Exclusion criteria for all groups included the presence of any known brain lesion or neurological condition (e.g. cerebral palsy, Tuberous Sclerosis, intraventricular hemorrhage of prematurity), the presence of any "hard" neurological sign (e.g., ataxia), orofacial abnormalities (e.g., cleft palate), bilinguality, severe intelligibility impairment, gross sensory or motor impairments, or identified mental retardation. For the TD group, exclusion criteria included, in addition, a history of psychiatric disturbance (e.g., ADHD, Anxiety Disorder), and having a family member with ASD or Developmental Language Disorder.

Methods

The present paper considers the results of three stress-related tasks: the Lexical Stress task, the Emphatic Stress task, and the Focus task (a modified version of the same task used in the PEPS-C). The Lexical Stress task (modified from Paul, Augustyn, Klin, & Volkmar, 2005) is a vocal imitation task where the computer plays a recording of a two-syllable nonsense word such as *tauveeb*, and the child repeats after the recorded voice with the same stress pattern. (Note that in the Paul et al. paradigm, the participant read a sentence aloud in which lexical stress -- e.g., *pre'sent* vs. *'present* -- had to be inferred from the sentential context; we also note that the English language contains only a few word pairs that differ in stress only, and that, moreover, as in the case of *present*, these words are not generally familiar to young children.) The Emphatic Stress task (Shriberg et al., 2001; Shriberg, Ballard, Tomblin, Duffy, Odell, & Williams, 2006) is also a vocal imitation task. Here, the child repeats an utterance in which one word is emphasized (*BOB may go home*, *Bob MAY go home*, etc.). Finally, in the Focus task (adapted from the PEPS-C), a recorded voice incorrectly describes a picture of a brightly colored animal, using the wrong word either for the animal or for the color. The child must correct the voice by putting contrastive stress on the incorrect word. These three tasks were selected because they span a wide range of prosodic capabilities (from vocal imitation tasks to generation of pragmatic prosody in a

picture description task), yet are strictly comparable in terms of their output requirements (i.e., putting stress on a word or syllable) and hence in terms of the dependent variables.

Previous results obtained on these tasks are few. They include poorer performance in ASD compared to TD on the Focus task (Peppé et al., 2007), and a significant difference on a lexical stress task (Paul et al., 2005; as noted, the latter was not a vocal imitation task and may thus make cognitive and verbal demands that differ from those made by our lexical stress task). The results from the Paul et al. (2005) study are difficult to relate to the present study because no group matching information was provided, tasks involved reading text aloud, the age range was broad and older (M age for ASD = 16.8, SD = 6.6), and certain tasks had serious ceiling effects.

The Focus and Lexical Stress tasks were preceded by their receptive counterparts, thereby establishing a basis for fully understanding these tasks. In the receptive Focus task, the child listened to pairs of pre-recorded sentences such as *I wanted CHOCOLATE and honey* or *I wanted chocolate and HONEY*. The child is asked to decide which food the speaker did not receive. In the receptive Lexical Stress task, the child listened to spoken names and had to judge whether the names were mispronounced; these names were high-frequency (e.g., *Robert, Nicole*), and were pronounced with correct (e.g., '*Robert*') or incorrect (e.g., '*Denise*') stress. These two tasks also started with four training trials during which the examiner corrected and, if necessary, modeled the response. Thus, significant efforts were made to ensure that the child understood the task requirements.

We note that in the Focus task no further modeling was provided after the training trials, so that this task cannot be considered a vocal imitation task in the same sense as the Emphatic Stress and Lexical Stress tasks; in addition, the preceding receptive Focus task – like all receptive tasks in the study – used a wide range of different voices, thereby avoiding presentation of a well-defined model that the child could imitate.

As mentioned in the Introduction, the recordings for a given child and task formed prosodic minimal pairs, such as '*tauveeb*' and '*tau'veeb*' in the Lexical Stress task. The child was generally not aware of this fact because the items occurred in a quasi-randomized order in which one member of a minimal pair rarely immediately followed the other member.

In cases where there was uncertainty about the (correct vs. incorrect) scoring judgments made during the examination, the examiner could optionally re-assess the scores after task completion by reviewing the audio recordings of the child's responses. We call the final scores resulting from this process "Examiner scores." We note that the examiners were blind to the child's diagnostic status only in the limited sense that they did not know the diagnostic group the child was assigned to and only administered the prosody tasks to the child (and not, e.g., the ADOS); however, in the course of administering even these limited tasks, an examiner unavoidably builds up an overall diagnostic impression that might influence his or her scores.

Analysis methods

The analysis methods are based on a general understanding of the acoustic patterns associated with stress and focus. A large body of phonetics research supports the following qualitative characterizations. When one compares the prosodic features extracted from a minimal pair such as '*tauveeb*' and '*tau'veeb*', the following pattern of acoustic differences is typically observed. First, in '*tauveeb*', F_0 rises at the start of the /tau/ syllable, reaches a peak in the syllable nucleus (/au/) or in the intervocalic consonant (/v/), and decreases in course of the second syllable; in '*tau'veeb*', F_0 shows little movement in the /tau/ syllable, rises at the start of the /veeb/ syllable, reaches a peak in the syllable nucleus (/ee/), and decreases again

(e.g., Caspers & van Heuven, 1993; van Santen & Möbius, 2000). Second, the amplitude contour shows a complicated, irregular pattern that reflects not only the prosodic contrast but also the phonetic segments involved (e.g., the /au/ vowel is louder than the /ee/ vowel, /t/ is louder than /v/, and vowels are generally louder than consonants). These segmental effects are stronger for amplitude than for F_0 , although also here segmental effects are considerable (e.g., Silverman, 1987; van Santen & Hirschberg, 1994). Yet, when one compares the same region of the same syllable in stressed and unstressed contexts (e.g., the initial part of the vowel of /tau/ in *'tauveeb* vs. *tau'veeb*), this region will have greater amplitude in the stressed case (e.g., van Santen & Niu, 2002; Miao et al., 2006). Third, the duration of the same syllable will be longer in stressed than the unstressed contexts (e.g., Klatt, 1976; van Santen, 1992; van Santen & Shih, 2000). Note, however, that even if these qualitative patterns of stress contrasts are typical, their quantitative acoustic manifestations vary substantially within and between speakers. This variability is a considerable challenge for instrumental methods.

It follows that for meaningful measurement of stress or focus a method is needed that (i) captures the dynamic patterns described above; (ii) factors out any segmental effects on F_0 or on amplitude; (iii) captures F_0 or amplitude patterns without being influenced by durational factors; and (iv) is insensitive to within- and between-speaker variability on dimensions that are not relevant for these patterns. These considerations led to a method called the “dynamic difference method” (van Santen et al., 2009), which we illustrate here using F_0 ; however, the method is equally applicable to other continuous features such as spectral balance or amplitude. This method critically relies on usage of prosodic minimal pairs to reduce the influence of segmental effects.

Figure 1 illustrates the successive steps of the algorithm. Based on our qualitative characterization of stress and focus, we expect the F_0 patterns of the members of a minimal pair that differ in the to-be-stressed syllable (in the Lexical Stress task) or word (in the Emphatic Stress and Focus tasks) to share the same shape (an up-down, single-peaked pattern) but have different peak locations; as a result, when the right-aligned pattern (i.e., having a later peak, as in *tau'veeb*) is subtracted from the left-aligned pattern (i.e., having an earlier peak, as in *'tauveeb*), after time-warping the two utterances so that the phonemes are aligned, the resulting difference curve generally shows an “up-down-up” pattern. This will be the case regardless of whether the utterances in a minimal pair are produced with different overall pitch, amplitude, or speaking rate. The measure hence is relatively insensitive to the child moving closer to the microphone, raising his voice, erratically changing his pitch range, or speeding up. The Figure shows these steps for two minimal pairs, one where the child produces a clear and correct prosodic distinction, and the other where the child does not. In the first minimal pair, the difference curve exhibits a clear up-down-up pattern, but it does not in the second pair. The computational challenge faced was how to quantitatively capture this distinction between the two pairs. Toward this end, van Santen et al. (2009) proposed an “up-down-up pattern detector” that measures the presence and strength of this pattern, using isotonic regression (Barlow, Bartholomew, Bremner, & Brunk, 1972). Specifically, the method approximates the difference curves with two ordinally constrained curves, one constrained to have an up-down-up pattern (“UDU curve”) and the other constrained to have a down-up-down pattern (“DUD curve”). If the difference curve exhibits a strong up-down-up pattern, as in the first example in the Figure, the UDU curve has a much better fit than the DUD curve; but when the difference curve exhibits a weak or no up-down-up pattern, the UDU and DUD curves fit about equally poorly. Van Santen et al. (2009) proposed a “dynamic difference measure” that compares the goodness-of-fits of the UDU and DUD curves to a given difference curve; it is normalized to range between +1 (for a correct contrast) and -1 (for an incorrect response); a value of 0 defines the boundary between correct and incorrect responses. The same measure is used for

amplitude. For duration we used as dynamic difference measure the quantity $(L_1R_2 - L_2R_1) / (L_1R_2 + L_2R_1)$, where L_i and R_i denote the duration of the i -th syllable or word in the left (L) and right (R) aligned items, respectively. Also the latter measure has the property of having larger, and positive, values when the correct distinction is made (i.e., by the child lengthening the stressed syllable or emphasized word), and having a theoretical range of -1 to $+1$.

To demonstrate that these dynamic difference measures can predict the judgments based on auditory-perceptual methods, the following procedure was used by van Santen et al. (2009). Six listeners independently judged the direction and the strength of contrast in prosodic minimal pairs. These judgments were truly “blind”, since the listeners had no information about, or direct contact with, the children, and listened to the minimal pairs in a random order, with each recording coming from a different child – thereby making it difficult to build up any overall impression of an individual child that could systematically bias the per-item judgments. We therefore consider the averages of the ratings of the six listeners (“Listener scores”) as the “gold standard”. To show that the dynamic difference measures can predict these Listener scores, a multiple regression analysis was performed in which regression parameters were estimated on a subset of the data and evaluated on the complementary set. Results indicated that the dynamic difference measures correlated with the Listener scores approximately as well as the individual listener ratings and better than the Examiner scores. Results also indicated that the F_0 dynamic difference measure was the strongest predictor of the Listener scores.

We will use the term “Simulated Listener scores” to refer to *composite* instrumental measures that are computed by applying the regression weights estimated in van Santen et al. (2009) to the dynamic difference measures computed from the children’s utterances in the present study. In other words, these Simulated Listener scores are our best guess at what the average scores of the six judges would have been; again, we contrast these scores with the Examiner scores, which are the scores assigned by the examiner during test administration and verified off-line after the session.

Results

The panels in Table II contain, for each of the three tasks, the group means and standard deviations, effect size (Cohen’s d , measured as the ratio of the difference between the means and the pooled standard deviation), t -test statistic, and the (two-tailed) p -value. The table shows that according to the Examiner scores the groups differ significantly on the Emphatic Stress task, differ with marginal significance on the Focus task, and do not differ on the Lexical Stress task. Second, according to the instrumental measures, the groups do not differ on the Emphatic Stress task using any of the individual dynamic difference measures or the Simulated Listener scores; do differ significantly on the Focus task using the Simulated Listener score; and also differ (but with ASD scoring *higher* than TD) on the Lexical Stress task using the F_0 dynamic difference measure and the Simulated Listener score. Note that in the latter task, this unexpected reversal is due primarily to the far greater value of the F_0 dynamic difference measure in the ASD group ($M = 0.534$) than in the TD group ($M = 0.299$); however, the ASD group also had slightly larger values than the TD group on the Amplitude and Duration measures, further contributing to the between-group difference on the Simulated Listener score since in the computation of this score all three dynamic difference measures have positive weights.

Linear Discriminant Analysis (LDA), with the F_0 , Amplitude, and Duration measures as predictors, and the diagnostic groups (TD vs. ASD) as the classes, confirmed the above picture, with no difference on the Emphatic Stress task ($F(2,42) = 0.318$, ns ; $d = 0.241$; note

that effect size is inflated when applied to LDA variates compared to one-dimensional variables because of the extra degrees of freedom in the form of the LDA weights), a significant difference on the Focus task ($F(2,42) = 4.019, p < 0.025; d = 0.819$), and a marginal difference on the Lexical Stress task ($F(2,42) = 2.525, p < 0.10; d = 0.703$).

Thus, our first hypothesis, which states that the instrumental methods can differentiate between the groups, is clearly confirmed by these results.

Both the Examiner and the instrumental measures broadly confirm the second hypothesis, which states that results will be task-dependent. However, the Examiner scores are not consistent with the specific hypothesis that the TD-ASD differential should be smaller on vocal imitation tasks (the Emphatic Stress and Lexical Stress tasks) than on non-imitation tasks, namely on the Focus task. The Emphatic Stress and Focus tasks, show, if anything, a reverse trend, with TD/ASD Examiner score ratios of 1.19 and 1.12 for the Emphatic Stress and Focus tasks, respectively, and corresponding effect sizes of 1.033 and 0.557.

On the other hand, the results from the instrumental measures support the second hypothesis. Particularly powerful in this respect is the better performance of the ASD group on the Lexical Stress task. A planned linear contrast, using the Simulated Listener scores and comparing the groups in terms of their differences between the vocal imitation tasks (the Emphatic Stress and Lexical Stress tasks) and the non-vocal imitation task (the Focus task) was significantly larger in the ASD group than in the TD group ($t(44) = 1.71, p < 0.05$, one-tailed).

We now turn to the third hypothesis. For both the Focus and the Lexical Stress task, the LDA weight for F_0 was negative (indicating that larger values on this measure were associated with the ASD group) while the weight was positive for duration (indicating that larger values on this measure were associated with the TD group). The Examiner and Simulated Listener scores, however, correlated positively with both the F_0 and the duration dynamic difference measures, regardless of whether these correlations were computed within the TD and ASD groups separately or in the pooled group. (All relevant correlations were significant at 0.05.) This indicates that the auditory-perceptual ratings, whether in the form of the Examiner scores or the Simulated Listener scores, predictably focused on the strength of the contrast as captured by the (positively weighted) sum of the F_0 and Duration measures. However, the negative LDA weight for the F_0 dynamic difference measure indicates that the groups differ primarily in the weighted *difference* between the F_0 and Duration measures, and not in their sum.

This suggests that the TD-ASD difference does not reside in weakness of the stress contrast but in an *atypical balance of the acoustic features*. We performed additional analyses of the Focus task to further investigate this interpretation.

In Figure 2, the participants are represented in an F_0 -by-Duration dynamic difference measure space. The scatter shows clear correlations between the two measures, suggesting an underlying stress strength factor. These correlations are significant for the ASD group ($r = 0.41, p < 0.025$), the TD group ($r = 0.68, p < 0.001$), and the two groups combined ($r = 0.56, p < 0.001$). The arrows indicate the directions in the plane that correspond to the Examiner and Simulated Listener scores, respectively. They point in the same (positive) direction as the major axis of the scatter, and are correlated positively with each other, with the two measures, and with the major axis (as measured via the first component, resulting from a Principal Components Analysis).

The tacit assumption of the auditory-perceptual method is that, if there is indeed a difference between the two groups, this difference entails weaker or less consistent expression of focus

by the ASD group; therefore, the ASD group should occupy a region in the scatter closer to the origin (the (0,0) point where the horizontal and vertical lines intersect). However, this is not how the groups differ. An indication of this is the substantial number (8 out of 23) of children in the ASD group who have slightly negative values on the Duration measure; no children in the TD group have negative values on this measure. This is not the case for the F_0 measure, where the corresponding numbers are 3 and 1, indicating that the duration results are unlikely to be caused by an inability of children in the ASD group to understand the task. This fundamental asymmetry of the groups in terms of the two measures can be further visualized by considering the location of the line that best separates the groups. The line correctly classifies 77% of the children into the correct group. Importantly, instead of separating the groups in terms of proximity to the origin, the line separates the groups in terms of the relative magnitudes of the two measures. This further confirms that what differentiates the groups is not the (weighted) sum of the two measures but their (weighted) difference. The arrow that is perpendicular to this line shows the direction in which the two groups maximally differ, and is, in fact, almost *orthogonal* to the arrows representing the auditory-perceptual scores.

We conducted a simple Monte Carlo simulation to statistically test these observations, again using the Focus task data. We fitted the pooled data using linear regression, resulting in a line similar to the best-separating line in Figure 2, measured the vertical (i.e., F_0) deviations of each data point from the regression line, and computed the standard two-sample *t*-test statistic, yielding a value of 3.11, reflecting that, as in Figure 2, the data points for children in the TD group were generally above the line and the data points for children in the ASD group below the line. We randomized group assignment 100,000 times, and found that the obtained value of 3.11 was exceeded only in 150 cases, yielding a *p*-value of 0.0015.

While the Focus task provides the clearest evidence for the difference between the groups in how they balance F_0 and duration, there is also a trend in the Lexical Stress task that supports this conclusion. The ratio of the F_0 and the duration dynamic difference measures is smaller in the TD group than in the ASD group ($t(44)=2.43$, $p<0.01$); the same is the case for the Focus task ($t(44)=1.81$, $p<0.05$). Although this effect was not statistically significant in the Emphatic Stress task, in all three tasks the ratio is smaller in the TD group (2.58, 1.71, and 3.79 for the Emphatic Stress, Focus, and Lexical Stress tasks, respectively) than in the ASD group (2.88, 2.24, and 6.21).

In combination, these results strongly support the third hypothesis. Unlike the auditory-perceptual scores, the key distinction between the TD and ASD groups does not reside in the overall strength with which prosodic contrasts are expressed but in a different balance of the degrees to which durational features and pitch features are used to express stress.

Discussion

The primary goal of this paper is to present results obtained on three prosodic tasks using new instrumental methods for the analysis of expressive prosody. The data analyzed were collected on study samples that were well-characterized, matched on key variables, of adequate size, and with a reasonably narrow age range. As mentioned in the introduction, few studies currently exist that meet these methodological criteria.

The data generally confirmed our hypotheses. First, the proposed instrumental methods were able to differentiate between ASD and TD groups in two of the three tasks. To our knowledge, this is the first time that such a finding has been reported for acoustic analyses that are fully automatic and capture dynamic prosodic patterns (such as the temporal

alignment of pitch movement with words or syllables) rather than global features such as average pitch.

Second, as predicted, the instrumental methods, and to a lesser degree the auditory-perceptual methods (in this study, the Examiner scores), showed weaker differences on the two vocal imitation tasks than on the pragmatic, picture-description task. In fact, on the Lexical Stress task, the instrumental methods showed better performance for the ASD group.

Third, as predicted by our third hypothesis, the instrumental methods revealed that the acoustic features that predict auditory-perceptual judgment are not the same as those that differentiate between the groups. Specifically, on the Focus task, auditory-perceptual judgment – both the Examiner scores and the Listener scores – is dominated by F_0 ; however, the key difference between the two groups was not in the use of F_0 but in the use of duration. This implies that the shortcoming of auditory-perceptual methods may be not only their unreliability and bias, but also their lack of attunement to what sets prosody apart in children with ASD (thereby possibly overlooking prosodic markers of ASD). One could imagine changing the auditory-perceptual task to judging prosodic atypicality, but we strongly doubt that this could be done reliably and in a truly blind manner.

There is, however, a deeper advantage of instrumental methods over auditory-perceptual methods. By providing a *direct acoustic profile* of speech rather than one mediated by subjective perception, results from instrumental methods can be linked – more directly than results based on perceptual ratings – to articulation, hence to the speech production process, and ultimately to underlying brain function. To illustrate this point, consider again the reported ASD-TD difference in the balance between pitch and duration in the Focus task. There is considerable evidence for lateralization of temporal and pitch processing in speech perception (for a recent review, see Zatorre & Gandour, 2008). Some studies indicate that temporal processing of auditory (including speech) input may be impaired in ASD (e.g., Cardy, Flagg, Roberts, Brian, & Roberts, 2004; Groen, van Orsouw, ter Huurne, Swinkels, van der Gaag, Buitelaar, et al., 2009), while pitch sensitivity may be a relative strength (e.g., Bonnel, Mottron, Peretz, Trudel, Gallun, E., & Bonnel, 2003). It is possible that inter-hemispheric underconnectivity in ASD (e.g., Just, Cherkassky, Keller, & Minshew, 2004) plays an additional role, resulting in poor coordination of pitch and temporal processing, thereby undermining any indirect benefit bestowed on temporal processing by intact pitch processing.

Unfortunately, few if any of these and other studies focus on speech production, and hence the link to our results is tentative at best. However, we may now have tools that provide precise quantitative measures of prosody production on a fine temporal scale (i.e., on a per-utterance instead of per-session, or per-individual, basis), which can then be used in combination with EEG, newer fMRI methods (e.g., Dogil, Ackermann, Grodd, Haider, Kamp, Mayer, et al., 2002), and magnetoencephalography to provide an integrated behavioral/neuroimaging account of expressive prosody in ASD that complements the above research on speech perception in ASD.

We emphasize that, while the Linear Discriminant analysis methods generated promising results (with upward of 75% correct classification for certain measures), it would be richly premature to use these measures for diagnostic purposes. This is not only the case because 75% is not adequate, but also because the profound heterogeneity of ASD makes it unlikely that any single or even small group of markers can serve these purposes. In addition, without additional groups in the study, such as children with other neurodevelopmental disorders or psychiatric conditions, the specificity of the proposed markers cannot be adequately

addressed. We note, however, that the purpose of this study was not to propose a diagnostic tool but to demonstrate the existence of innovative prosody-based markers of ASD that critically rely on computation and whose discovery would have been unlikely if one were to solely rely on conventional auditory-perceptual methods.

As a final note, our results should be considered preliminary because they are the first ones reported of this general nature. They thus clearly require confirmation in subsequent studies, including studies with larger samples and different group matching strategies.

Acknowledgments

We thank Sue Peppé for making available the pictorial stimuli from the PEPS-C and for granting us permission to create new versions of several of the PEPS-C tasks; Rhea Paul for suggesting the Lexical Stress task, and for permitting modifications to it; the clinical staff at OHSU (Beth Langhorst, Rachel Coulston, and Robbyn Sanger Hahn) and at Yale University (Nancy Fredine, Moira Lewis, Allyson Lee) for data collection; senior programmer Jacques de Villiers for the data collection software and data management architecture; and especially the parents and children who participated in the study. This research was supported by a grant from the National Institute on Deafness and Other Communicative Disorders, NIDCD 1R01DC007129-01 (van Santen, PI); a grant from the National Science Foundation, IIS-0205731 (van Santen, PI); by a Student Fellowship from AutismSpeaks to Emily Tucker Prud'hommeaux; and by an Innovative Technology for Autism grant from AutismSpeaks (Roark, PI). The views herein are those of the authors and reflect the views neither of the funding agencies nor of any of the individuals acknowledged.

References

- American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Third Edition (DSM-III). Washington, DC: American Psychiatric Association; 1980.
- American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised (DSM-III-R). Washington, DC: American Psychiatric Association; 1987.
- American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR). Washington, DC: American Psychiatric Association; 2000.
- Baltaxe, C. Acoustic characteristics of prosody in autism. In: Mittler, P., editor. *Frontier of knowledge in mental retardation*. Baltimore, MD: University Park Press; 1981. p. 223-233.
- Baltaxe C. Use of contrastive stress in normal, aphasic, and autistic children. *Journal of Speech and Hearing Research*. 1984; 27:97–105. [PubMed: 6201678]
- Baltaxe, C.; Simmons, J.; Zee, E. Intonation patterns in normal, autistic and aphasic children. In: Cohen, A.; van de Broecke, M., editors. *Proceedings of the Tenth International Congress of Phonetic Sciences*. Dordrecht, The Netherlands: Foris Publications; 1984. p. 713-718.
- Barlow, R.; Bartholomew, D.; Bremner, J.; Brunk, H. *Statistical Inference Under Order Restrictions*. New York: Wiley; 1972.
- Baron-Cohen, S. Theory of mind and autism: A review. In: Glidden, LM., editor. *International Review of Research in Mental Retardation* vol.23: Autism. New York: Academic Press; 2000.
- Beach C. The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence of cue trading relations. *Journal of Memory and Language*. 1991; 30:644–663.
- Berument S, Rutter M, Lord C, Pickles A, Bailey A. A autism screening questionnaire: diagnostic validity. *The British Journal of Psychiatry*. 1999; 175:444–451. [PubMed: 10789276]
- Bonnel A, Mottron L, Peretz I, Trudel M, Gallun E, Bonnel AM. Enhanced pitch sensitivity in individuals with autism: A signal detection analysis. *Journal of Cognitive Neuroscience*. 2003; 15:226–235. [PubMed: 12676060]
- Campbell, N.; Beckman, M. Stress, prominence and spectral balance. In: Botinis, A.; Kouroupetroglu, G.; Carayiannis, G., editors. *Intonation: Theory, Models and Applications*. Athens, Greece: ESCA and University of Athens; 1997. p. 67-70.
- Cardy J, Flagg E, Roberts W, Brian J, Roberts T. Magnetoencephalography identifies rapid temporal processing deficit in autism and language impairment. *NeuroReport*. 2005; 16:329–332. [PubMed: 15729132]

- Caspers J, van Heuven V. Effects of time pressure on the phonetic realisation of the Dutch accent-lending pitch rise and fall. *Phonetica*. 1993; 50:161–171. [PubMed: 8295924]
- Dawson G, Webb S, Schellenberg G, Dager S, Friedman S, Aylward E, Richards T. Defining the broader phenotype of autism: Genetic, brain, and behavioral perspectives. *Development and Psychopathology*. 2002; 14:581–611. [PubMed: 12349875]
- Diehl J, Watson D, Bennetto L, McDonough J, Gunlogson C. An acoustic analysis of prosody in high-functioning autism. *Applied Psycholinguistics*. 2009; 30:385–404.
- Dogil G, Ackermann H, Grodd W, Haider H, Kamp H, Mayer J, Riecker A, Wildgruber D. The speaking brain: a tutorial introduction to fMRI experiments in the production of speech, prosody and syntax. *Journal of Neurolinguistics*. 2002; 15:59–90.
- Fine J, Bartolucci G, Ginsberg G, Szatmari P. The use of intonation to communicate in pervasive developmental disorders. *Journal of Child Psychology and Psychiatry*. 1991; 32:771–782. [PubMed: 1918227]
- Fosnot, S.; Jun, S. Prosodic characteristics in children with stuttering or autism during reading and imitation. In: Ohala, JJ.; Hasegawa, Y., editors. *Proceedings of the 14th International Congress of Phonetic Sciences*. 1999. p. 1925–1928.
- Fry D. Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*. 1955; 27:765–768.
- Gillberg, C.; Ehlers, S. High functioning people with autism and Asperger's syndrome. In: Schopler, E.; Mesibov, G.; Kunc, L., editors. *Asperger Syndrome or High Functioning Autism?*. New York: Plenum Press; 1998. p. 79–106.
- 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]
- Gotham K, Risi S, Dawson G, Tager-Flusberg H, Joseph R, Carter A, Hepburn S, McMahon W, Rodier P, Hyman S, Sigman M, Rogers S, Landa R, Spence A, Osann K, Flodman P, Volkmar F, Hollander E, Buxbaum J, Pickles A, Lord C. A Replication of the Autism Diagnostic Observation Schedule (ADOS) Revised Algorithms. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2008; 47:642–651. [PubMed: 18434924]
- Groen W, van Orsouw L, ter Huurne N, Swinkels S, van der Gaag R, Buitelaar J, Zwier M. Intact Spectral but Abnormal Temporal Processing of Auditory Stimuli in Autism. *Journal of Autism and Developmental Disorders*. 2009; 39:742–750. [PubMed: 19148738]
- Jarrold C, Brock J. To match or not to match? Methodological issues in autism-related research. *Journal of Autism and Developmental Disorders*. 2004; 34:81–86. [PubMed: 15098961]
- Joseph RM, Tager-Flusberg H, Lord C. Cognitive profiles and social-communicative functioning in children with autism spectrum disorder. *Journal of Child Psychology and Psychiatry*. 2002; 43:807–821. [PubMed: 12236615]
- Just M, Cherkassky V, Keller T, Minshew N. Cortical activation and synchronization during sentence comprehension in high-functioning autism: Evidence of underconnectivity. *Brain*. 2004; 127:1811–1821. [PubMed: 15215213]
- Kanner L. Autistic disturbances of affective content. *Nervous Child*. 1943; 2:217–250.
- Kent R. Hearing and Believing: Some Limits to the Auditory-Perceptual Assessment of Speech and Voice Disorders. *American Journal of Speech-Language Pathology*. 1996; 5:7–23.
- Klatt D. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*. 1976; 59:1208–1221. [PubMed: 956516]
- Kreiman J, Gerratt B. Validity of rating scale measures of voice quality. *Journal of the Acoustical Society of America*. 1998; 104:1598–1608. [PubMed: 9745743]
- Leyfer OT, Tager-Flusberg H, Dowd M, Tomblin JB, Folstein SE. Overlap between autism and specific language impairment: Comparison of autism diagnostic interview and autism diagnostic observation schedule scores. *Autism Research*. 2008; 1:284–296. [PubMed: 19360680]
- Lord C, Risi S, Lambrecht L, Cook E, Leventhal B, DiLavore P, Pickles A, Rutter M. 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*. 2000; 30:205–223. [PubMed: 11055457]

- McCann J, Peppé S. Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders*. 2003; 38:325–50. [PubMed: 14578051]
- McCann J, Peppé S, Gibbon F, O'Hare A, Rutherford M. Prosody and its relationship to language in school-aged children with high-functioning autism. *International Journal of Language & Communication Disorders*. 2007; 42:682–702. [PubMed: 17885824]
- Miao, Q.; Niu, X.; Klabbers, E.; van Santen, J. Proceedings of the Third International Conference on Speech Prosody. Germany: Dresden; 2006. Effects of prosodic factors on spectral balance: Analysis and synthesis.
- Paul R, Bianchi N, Augustyn A, Klin A, Volkmar F. Production of Syllable Stress in Speakers with Autism Spectrum Disorders. *Research in Autism Spectrum Disorders*. 2008; 2:110–124. [PubMed: 19337577]
- Paul R, Augustyn A, Klin A, Volkmar F. Perception and production of prosody by speakers with autism spectrum disorders. *Journal of Autism and Developmental Disorders*. 2005; 35:201–220.
- Peppé S, McCann J. Assessing intonation and prosody in children with atypical language development: the PEPS-C test and the revised version. *Clinical Linguistics and Phonetics*. 2003; 17:345–354. [PubMed: 12945610]
- Peppé 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.
- Post, B.; d'Imperio, M.; Gussenhoven, C. Proceedings of the 16th International Congress of Phonetic Sciences. Saarbruecken; 2007. Fine phonetic detail and intonational meaning; p. 191-196.
- Sabbagh M. Communicative intentions and language: evidence from right hemisphere damage and autism. *Brain and Language*. 1999; 70:29–69. [PubMed: 10534371]
- Shriberg L, Ballard K, Tomblin J, Duffy J, Odell K, Williams C. Speech, prosody, and voice characteristics of a mother and daughter with a 7;13 translocation affecting FOXP2. *Journal of Speech, Language, and Hearing Research*. 2006; 49:500–525.
- Shriberg L, Paul R, McSweeney J, Klin A, Cohen D, Volkmar F. Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger's Syndrome. *Journal of Speech, Language, and Hearing Research*. 2001; 44:1097–1115.
- Siegel D, Minshew N, Goldstein G. Wechsler IQ profiles in diagnosis of high-functioning autism. *Journal of Autism and Developmental Disorders*. 1996; 26:389–406. [PubMed: 8863091]
- Silverman, K. The Structure and Processing of Fundamental Frequency Contours. University of Cambridge (UK); 1987. Doctoral Thesis
- Sluijter, A.; Shattuck-Hufnagel, S.; Stevens, K.; van Heuven, V. Proceedings of the 13th International Congress of Phonetic Sciences. Stockholm; 1995. Supralaryngeal resonance and glottal pulse shape as correlates of stress and accent in English; p. 630-633.
- Sluijter A, van Heuven V, Pacilly J. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*. 1997; 101:503–513.
- Tager-Flusberg, H. Understanding the language and communicative impairments in autism. In: Glidden, LM., editor. *International Review of Research on Mental Retardation*. Vol. Volume 20. San Diego, CA: Academic Press; 2000. p. 185-205.
- van Santen J. Contextual effects on vowel duration. *Speech Communication*. 1992; 11:513–546.
- van Santen, J.; Hirschberg, J. Proceedings of the International Conference on Spoken Language Processing. Yokohama, Japan: 1994. Segmental effects on timing and height of pitch contours; p. 719-722.
- van Santen, J.; Möbius, B. A quantitative model of F0 generation and alignment. In: Botinis, A., editor. *Intonation - Analysis, Modeling and Technology*. Dordrecht: Kluwer; 2000. p. 269-288.
- van Santen J, Shih C. Suprasegmental and segmental timing models in Mandarin Chinese and American English. *Journal of the Acoustical Society of America*. 2000; 107:1012–1026. [PubMed: 10687710]
- van Santen, J.; Niu, X. Proceedings of the IEEE Workshop on Speech Synthesis. California: Santa Monica; 2002. Prediction and synthesis of prosodic effects on spectral balance.
- van Santen J, Prud'hommeaux ET, Black L. Automated Assessment of Prosody Production. *Speech Communication*. 2009; 51:1082–1097. [PubMed: 20160984]

Zatorre R, Gandour J. Neural specializations for speech and pitch: moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B*. 2008; 363:1087–1104.

Figure 1.

The dynamic difference method, illustrated for well-differentiated responses (panels a-e) and for undifferentiated responses (panels f-j). The F_0 contours for the left aligned (e.g., 'blue sheep') and right aligned (e.g., blue 'sheep') items are displayed in the top two rows (panels a, b, f, g). The third row (panels c and h) displays the same contours after time warping so that the phoneme boundaries coincide. The next row (panels d and i) shows the difference contour, obtained by subtracting the time-warped right-aligned contour from the left-aligned contour; the continuous thick curve is the best-fitting up-down-up curve, and the continuous thin curve the best-fitting down-up-down curve. Finally, the bottom row (panels e and j) shows the deviations of these respective continuous curves from the difference contour,

again with the thicker curve corresponding to the up-down-up fit and the thinner line to the down-up-down fit; it can be seen that the thicker curve displays a smaller overall deviation (in fact, almost no deviation) than the thinner curve in the well-differentiated case, but not in the undifferentiated case.

Figure 2.

Scatter plot of the values on the duration and F_0 difference measures of the ASD (+) and TD (O) groups in the Focus task. The thick line is the line that best separates the two groups. The dashed arrow represents the Simulated Listener scores, with the score of a given data point defined by its projection onto this arrow. The dotted arrow represents the Examiner scores. The thick solid arrow is perpendicular to the best-separating line and represents the dimension on which the two groups maximally differ. These data suggest that whereas listeners and examiners base their judgments on the weighted sum of the F_0 and duration difference measures, it appears that the two groups differ on the weighted difference between these measures.

Means (*M*) and standard deviations (*SD*) for the TD and ASD groups for age, non-verbal IQ (standardized scores), and subtests of the Wechsler Scales (scaled scores) on which non-verbal IQ is based. Significance values are one-tailed (*ns*: $p>0.10$).

Table 1

Measure	TD <i>M</i>	ASD <i>M</i>	TD <i>SD</i>	ASD <i>SD</i>	<i>p</i> (1-tailed)
Age	6.35	6.57	1.02	1.29	<i>ns</i>
NVIQ	119.92	117.63	8.57	11.48	<i>ns</i>
Block Design	12.81	13.96	3.10	2.46	<i>ns</i>
Matrix Reasoning	13.65	13.26	2.13	2.84	<i>ns</i>
Picture Concepts	13.15	11.37	2.09	2.79	0.01

Table II

Means (*MN*) and standard deviations (*SD*) for the TD and ASD groups for the Examiner scores and instrumental measures. Significance values are two-tailed (*ns*; $p > 0.10$).

Emphatic Stress Task						
Measure	TD <i>M</i>	ASD <i>M</i>	TD <i>SD</i>	ASD <i>SD</i>	Effect	<i>t</i> <i>p</i> (2 tailed)
Examiner	0.974	0.818	0.079	0.192	1.033	3.297 0.002
F ₀	0.582	0.527	0.266	0.299	0.195	0.615 ns
Amplitude	0.349	0.315	0.258	0.222	0.144	0.466 ns
Duration	0.226	0.183	0.135	0.174	0.268	0.857 ns
Simul. Listener	0.463	0.401	0.208	0.260	0.263	0.839 ns
Focus Task						
Examiner	0.847	0.759	0.132	0.179	0.557	1.803 0.08
F ₀	0.393	0.316	0.291	0.284	0.270	0.875 ns
Amplitude	0.178	0.102	0.173	0.301	0.306	0.990 ns
Duration	0.230	0.141	0.179	0.134	0.570	1.800 0.08
Simul. Listener	0.435	0.256	0.292	0.274	0.633	2.028 0.05
Lexical Stress Task						
Examiner	0.956	0.919	0.070	0.103	0.408	1.269 ns
F ₀	0.299	0.534	0.310	0.270	-0.809	-2.588 0.015
Amplitude	0.203	0.260	0.245	0.180	-0.264	-0.835 ns
Duration	0.079	0.086	0.080	0.104	-0.080	-0.250 ns
Simul. Listener	0.309	0.460	0.232	0.222	-0.665	-2.127 0.05