Published in final edited form as:
Autism Res. 2017 December ; 10(12): 1991–2001. doi:10.1002/aur.1847.

Production and perception of emotional prosody by adults with autism spectrum disorder

Daniel J. Hubbard, **Daniel J. Faso**, **Peter F. Assmann**, and **Noah J. Sasson**
School of Behavioral and Brain Sciences, University of Texas at Dallas, GR41, 800 West, Campbell Road, Richardson, TX 75080, USA

Abstract

This study examined production and perception of affective prosody by adults with autism spectrum disorder (ASD). Previous research has reported increased pitch variability in talkers with ASD compared to typically-developing (TD) controls in grammatical speaking tasks (e.g., comparing interrogative vs. declarative sentences), but it is unclear whether this pattern extends to emotional speech. In this study, speech recordings in five emotion contexts (angry, happy, interested, sad, and neutral) were obtained from 15 adult males with ASD and 15 controls (Experiment 1), and were later presented to 52 listeners (22 with ASD) who were asked to identify the emotion expressed and rate the level of naturalness of the emotion in each recording (Experiment 2). Compared to the TD group, talkers with ASD produced phrases with greater intensity, longer durations, and increased pitch range for all emotions except neutral, suggesting that their greater pitch variability was specific to emotional contexts. When asked to identify emotion from speech, both groups of listeners were more accurate at identifying the emotion context from speech produced by ASD speakers compared to TD speakers, but rated ASD emotional speech as sounding less natural. Collectively, these results reveal differences in emotional speech production in talkers with ASD that provide an acoustic basis for reported perceptions of oddness in the speech presentation of adults with ASD.

Lay Summary—This study examined emotional speech communication produced and perceived by adults with autism spectrum disorder (ASD) and typically-developing (TD) controls. Compared to the TD group, talkers with ASD produced emotional phrases that were louder, longer, and more variable in pitch. Both ASD and TD listeners were more accurate at identifying emotion in speech produced by ASD speakers compared to TD speakers, but rated ASD emotional speech as sounding less natural.

Keywords

Autism spectrum disorder; emotion; affective prosody; expressive speech; vocal affect; speech production; speech perception

Corresponding author: Daniel J. Hubbard School of Behavioral and Brain Sciences The University of Texas at Dallas, GR41 800 West Campbell Road Richardson, TX 75080, dhubbard@utdallas.edu Phone: 469-371-5831.

Introduction

Affective prosody, defined as the use of nonlinguistic features in speech to convey emotion, contributes to effective communication and social functioning (Banse & Scherer, 1996; Scherer, 2003). Affective prosody production in typically developing (TD) populations contains systematic differences in acoustic properties associated with discrete emotion categories (Scherer, 1979, 1986; Murray & Arnott, 1993; Juslin & Laukka, 2003), including modulations in fundamental frequency (f0; associated with perceived voice pitch), intensity (e.g., loudness), and duration, which are correlated with a general state of physiological arousal (Banse & Scherer, 1996). High arousal emotions such as happiness are characterized by increases in f0 range, intensity and duration, whereas decreased f0 range, intensity and duration are typical of low arousal emotions such as sadness (Scherer, 1979, 1986; Murray & Arnott, 1993). For example, anger is often conveyed not only by linguistic content (e.g., what is said) but can also be reliably identified by nonlinguistic content (e.g., how it is said), including increased f0 variability, greater intensity and shorter durations compared to neutral utterances. Indeed, the perceived meaning of an utterance can be altered based on modulations to emotion-relevant acoustic properties. Even in short utterances such as the word “hello,” the talker’s intended meaning can vary to convey happiness (rising and falling pitch contour, faster speech rate) or sadness (falling pitch contour, slower speech rate), or a range of other emotions dependent upon reliable changes in relevant acoustic properties.

For individuals with autism spectrum disorder (ASD), pervasive social disability manifests not only in difficulties in social perception and cognition (Sasson et al., 2011), but also in differences in social expressivity (Begeer et al., 2008). Individuals with ASD are characterized by distinct social presentations, including atypical facial affect (Faso, Sasson & Pinkham, 2015) and distinct affective speech patterns (Fosnot & Jun, 1999; Nadig & Shaw, 2012) that begin in childhood and are associated with overall language ability (Lyons, Simmons, & Paul, 2014). Consequently, the social presentation of individuals with ASD is reliably rated as more odd or awkward by potential social partners (Van Bourgondien & Woods, 1992; Paul et al., 2005; Sasson et al., 2017). In turn, these judgments are associated with reduced inclinations on the part of peers to engage socially (Sasson et al., 2017), a process that may contribute to reduced quantity and quality of social experiences for those with ASD. However, the specific acoustic properties linked to perceptions of oddness in speech produced by individuals with ASD have not been identified (Nadig & Shaw, 2012).

The few studies that have analyzed acoustic properties of prosody in talkers with ASD have almost exclusively focused on grammatical prosody (used to signal syntactic information such as whether an utterance is a question or a statement) or pragmatic prosody (used for social information such as highlighting for the listener information that is new to a conversation). These studies have found that speech from individuals with ASD is characterized by abnormal sentence stress and intonation patterns, including using greater f0 range compared to control participants (Fosnot & Jun, 1999; Nadig & Shaw, 2012; Green & Tobin, 2009). Increased f0 range produced by speakers with ASD is inconsistent with traditional accounts of “flat” or “monotone” speaking styles in ASD, and parallels recent work showing exaggerated and more intense, not flat, facial displays in adults with ASD (Faso, Sasson, & Pinkham, 2015).

Whether emotional prosody similarly differs for individuals with ASD has not been well-studied. This is a surprising oversight, given that ASD is characterized by difficulties in the perception of emotion (Uljarevic & Hamilton, 2013), and differences in emotional expressivity can affect social functioning (Sasson et al., 2017). The one study that examined affective prosody in ASD using emotion elicitation (Hubbard & Trauner, 2007) found higher mean level f0 in children with ASD for sad and angry phrases, but these group differences did not reach significance. However, children in that study were encouraged to repeat voice manipulations used by a model talker, which may have affected their natural use of emotional prosody and contributed to the reported absence of group effects.

Hubbard and Trauner (2007) also did not investigate whether emotional prosody produced by speakers with ASD affects the ability of listeners to perceive the emotion being expressed. If emotional prosody is harder to detect in speakers with ASD, or is judged more awkward or less natural relative to controls, the quality of their social interaction and communication may be affected. In fact, although several studies have demonstrated that individuals with ASD are less adept at identifying emotion from speech prosody (Golan et al., 2007; Peppé & McCann, 2003; Rutherford et al., 2002; Wang et al., 2007), to our knowledge no study to date has examined whether emotional prosody produced by individuals with ASD is less effective at communicating emotional information to listeners.

The objectives of the current study were to determine whether acoustic differences in production of affective prosody exist between adult talkers with ASD and controls, and to examine whether perception of affective prosody by naïve listeners differs between the two talker groups. A set of five emotionally-ambiguous phrases (e.g., “I can’t believe this”) was obtained from talkers with ASD and TD controls using an evoked elicitation technique in five emotion contexts: angry, happy, interested, sad, and neutral. The recordings protocol was identical for each talker to enable detailed utterance-level group comparisons between talkers with ASD and controls. Based on prior work showing increased f0 range and longer utterance durations produced by talkers with ASD in grammatical and pragmatic prosody tasks (e.g., Fosnot & Jun, 1999; Nadig & Shaw, 2012), we predicted that phrases produced by talkers with ASD in affective tasks would exhibit increased f0 range and longer durations compared to those produced by controls.

A subset of the recordings was presented to a group of listeners with ASD as well as TD listeners in perceptual tests consisting of emotion context judgments and naturalness ratings. This resulted in a 2×2 experimental design in which ASD and TD listener groups responded to phrases produced by ASD and TD talker groups, enabling analysis of both group-level main effects and interactions between the talker and listener groups. We predicted that emotion recognition accuracy (ERA) and naturalness ratings would be lower for utterances produced by talkers with ASD compared to those produced by TD talkers.

Experiment 1: Affective Prosody Production

Participants

Fifteen adult males with an ASD diagnosis (mean age 27 years; age range 21–42 years) and fifteen TD males (mean age 21 years; age range 18–26 years) were recruited as talkers for this study. Talkers with ASD were recruited through the University of Texas at Dallas (UTD) Autism Research Collaborative (ARC), a database of adults with ASD in the local area who expressed interest in research participation. All participants with ASD in our study had prior DSM-IV or DSM-5 diagnoses, and were included in the study if they met or exceeded the revised ADOS 2 Module 4 cut-off (Hus & Lord, 2014) on the Autism Diagnostic Observation Schedule (Lord et al., 2000) conducted by a certified clinician in our lab group. Each talker's cognitive ability was assessed using the Wechsler Abbreviated Scales of Intelligence (WASI; Wechsler, 1999). Participants with ASD were provided with financial compensation for participating.

TD participants were recruited from the UTD Behavioral and Brain Sciences undergraduate participant pool and were awarded research credit as compensation. Each TD participant self-reported no diagnosis of autism and completed the Broad Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007) to assess the presence of traits associated with the Broad Autism Phenotype (BAP). Fourteen of fifteen TD talkers had an overall BAPQ score below the threshold used to classify the presence of the BAP (Sasson et al., 2013). Recordings produced by the talker who surpassed the increased specificity score were analyzed separately, and given that patterns of affective prosody production were consistent with other TD talkers, the recordings were included in the acoustic analysis and listening test. All participants spoke English as a first language and none had a detectable regional accent. The ASD group IQ mean was 107 (range = 88–130; *SD* = 12.69), compared to 115 (range = 101–129; *SD* = 9.36) for TD talkers. This difference did not reach significance ($p = .053$). The groups did differ on age ($F(1,28)=16.65$, $p<.001$). However, there were no significant effects of age on any of the dependent variables.

Method and Procedure

Recordings of affective prosody occurred during an evoked elicitation procedure similar to that used with adults with ASD in Faso, Sasson & Pinkham (2015). Talkers were asked to recall personal emotional experiences relevant to each of the five targeted emotion contexts: neutral, angry, happy, interested, and sad, then produce the phrases in the target emotion context while recalling the details of those experiences (see Faso, Sasson & Pinkham, 2015 for validation details of this method). For the neutral context, participants were asked to recall a time when they felt no particular emotion. The participants were given as much time as needed to recall a personal emotional experience and acclimate to the target emotion prior to beginning the recordings.

The five emotional contexts were selected because they occupy distinct regions along the arousal level and valence dimensions, which have long been viewed as important components of human emotional response (Wundt, 1909; Russell, 1980, 2003). Angry and happy contexts occupy opposite ends of the high arousal dimension (angry speech is high arousal, negative valence; happy speech is high arousal, positive valence); sad and interested contexts occupy opposite ends of the low arousal dimension (sad speech is low arousal, negative valence; interested speech is low arousal, positive valence). The collected speech consisted of five emotionally ambiguous phrases similar to those used in previous studies

(Hubbard & Trauner, 2007; Bänziger, Mortillaro, & Scherer, 2012), and each phrase was produced in each of the five emotion contexts.

Each context was recorded separately – beginning with neutral – followed by the other four contexts in random order. Neutral recordings were produced first to establish an unexpressive baseline for each talker and eliminate the possibility that a previously recorded context would influence neutral productions. Consistent with prior work in this area (e.g., Hubbard & Trauner, 2007), a controlled set of phrases was used so that analyses involved speech in which the lexical content was the same. The following phrases were chosen because each can be realistically produced in any of the five emotion contexts: 1) “What do you mean?” 2) “Why did you do that?” 3) “I can’t believe this,” 4) “Yes, that’s what I meant,” and 5) “Well, how do you know?” With the exception of the neutral context, talkers were asked to produce each phrase three times successively in increasingly expressive repetitions (e.g., happy, happier, happiest), which allowed us to determine if talkers with ASD differ from controls in their utilization of voice parameters when increasing expressivity. A printed copy of the phrases was provided to talkers during the elicitation task. The recordings were produced in a sound-attenuated booth, digitized with 16-bit resolution and stored at a 48 kHz sample rate. The talker instructions, audio/text prompts and data collection tasks were semi-automated using custom Matlab scripts. An experimenter accompanied each talker in the sound booth to assist with the recording, and no talkers reported having any difficulty completing the protocol.

Data Analysis

Each talker produced 75 evoked phrases, consisting of three repetitions of five phrases produced in five emotion contexts, for a total of 2,250 phrases from 30 talkers. Individual utterances were isolated by truncating leading and trailing silence and their acoustic properties were measured at 1-ms intervals. F0, the dominant voice cue important for perception of a talker’s affective state (Fairbanks & Provonost, 1939; Williams & Stevens, 1972) was estimated using STRAIGHT (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999). For the f0 estimation, segments of silence between words were removed to eliminate tracking errors. The vector of f0 measurements for each utterance was time-normalized by linear interpolation to a fixed length of 500 samples. Intensity (dB SPL) estimates were obtained at 1-ms intervals using Praat (Boersma & Weenink, 2014), and duration was calculated by taking the total time in ms from the onset of the first word to the offset of the final word in each phrase. Silent intervals were included in duration measurements because patterns of silence in a phrase may have relevance for perceptions of social oddness when communicating with a talker with ASD.

Statistical analysis of the acoustic measurements was performed using mixed effects analysis of variance (ANOVA) to identify properties that are reliable predictors for the separate emotion contexts and to determine if group differences in affective prosody production exist between ASD and TD talkers. Separate analyses were performed on f0 range (Hz), mean intensity (dB SPL) and duration (ms) for each talker group (ASD vs. TD). Within-group factors were emotional context (neutral, angry, happy, interested, & sad), phrase (five ambiguous phrases) and repetition (three repetitions). Talkers were treated as a random effect. Thus, comparisons were made between utterances in which ASD and TD talkers produced identical target syllables or phrases in the same elicited emotion contexts. Using this method, any differences found were reflective of divergent patterns of prosody production in the ASD and TD groups and are not the result of differences in linguistic content. The R programming language (R Core Team, 2015) and the “lme4” package (Bates et al., 2015) were used to perform the analysis.

Results

For each utterance, f0 range (Hz), mean intensity (dB SPL) and duration (ms) were analyzed to examine acoustic differences in affective prosody production between the ASD and TD groups. Table 1 lists the mean, standard deviation, and range for f0, intensity, and duration for evoked phrases for each emotion context and talker group.

The analysis of f0 range revealed main effects for talker group ($F(1,28)=4.65$, $p<.05$), emotion context ($F(4,2069)=110.50$, $p<.001$) and phrase ($F(4,2069)=13.548$, $p <.001$), and interactions between talker group and emotion context ($F(4,2069)=14.29$, $p<.001$), talker group and phrase ($F(4,2069)=2.45$, $p<.05$), and emotion context by phrase ($F(16,2069)=3.95$, $p<.001$). Planned comparisons for each emotion context confirmed the prediction that f0 range was greater in phrases produced by talkers with ASD for each emotion context except neutral. Figure 1 shows mean f0 range calculations by group and context. Phrases produced by talkers with ASD had a mean f0 range of 119 Hz (SD=91 Hz), compared to 93 Hz (SD=67 Hz) for TD talkers.

Individual f0 contours for the phrase “Why did you do that” are shown in Figure 2 (light grey lines), with 95% confidence bands superimposed on top (black shaded areas). The wider confidence bands for the ASD group suggest that patterns of f0 production in phrases spoken by talkers with ASD were more variable (left-side panels), compared to controls (right-side panels). Both groups produced neutral phrases with a relatively flat f0 contour characteristic of speech produced in a neutral tone. For the other emotion contexts, however, it was evident that the TD talkers followed a more consistent pattern of f0 production compared to talkers with ASD.

The analysis of mean intensity resulted in main effects for talker group ($F(1,28)=4.22$, $p<.05$), emotion context ($F(4,2072)=69.80$, $p<.001$) and phrase ($F(4,2072)=49.83$, $p<.001$). Significant interactions were found for talker group by emotion context ($F(4,2072)=10.72$, $p<.001$), talker group by phrase ($F(4,2072)=5.35$, $p<.001$), emotion context by phrase ($F(16,2072)=3.31$, $p<.001$), and emotion context by repetition ($F(8,2072)=8.91$, $p<.001$). The three-way interaction between talker group, emotion context and phrase was also significant ($F(16,2072)=2.21$, $p<.05$). The mean intensity for the ASD talker group was 66.27 dB SPL (SD=3.95 dB SPL) compared to 64.90 dB SPL (SD=3.66 dB SPL) for the TD talker group.

Bonferroni-corrected post-hoc tests revealed that sad and neutral phrases were produced by talkers with ASD with greater mean intensity than those produced by TD talkers. Phrases produced by talkers with ASD were up to 2.5 dB higher in voice intensity compared to TD talkers. Figure 3 shows mean intensity levels by talker group, emotion context and repetition. The pattern of results indicates that both groups of talkers used similar modulations to intensity to convey increased expressiveness, but that talkers with ASD produced the phrases with greater overall intensity. Figure 3 also shows how repetition interacted with emotional context, with intensity increasing most sharply with repetition for anger. Thus, prosody production patterns continued in the expected direction for both groups as talkers became more expressive with each repetition. Table 1 shows that intensity variation, as measured by the range and standard deviation of dB SPL across the entire phrase, was also higher for talkers with ASD compared to controls. One exception was the interested context, where no differences were found.

The analysis of phrase duration revealed main effects for emotion context ($F(4, 2057) = 46.26$, $p < .001$), phrase ($F(4, 2057) = 218.89$, $p < .001$), and repetition ($F(4,2057)=10.24$, $p<.001$), but the main effect of group did not reach significance ($p=.069$). There were significant interactions between talker group and emotion context ($F(4,2057)=24.23$, $p<.001$), talker group and phrase ($F(4,2057)=10.26$, $p<.001$), and emotion context and phrase ($F(16,2057)=3.35$, $p<.001$). The mean phrase duration for the talker group with ASD was 1242.04 ms (*SD*=352.88 ms), compared to 1124.98 ms (*SD*=283.84 ms) for the TD talker group. Bonferroni-corrected post-hoc tests were performed to compare mean durations for the two talker groups per emotion context, and revealed that interested phrases produced by talkers with ASD were longer than those produced by controls. The talker group comparisons for the happy ($p=.051$) and sad contexts ($p=.071$) approached significance. Figure 4 shows mean phrase durations by talker group, emotion context and repetition, and reveals that talkers with ASD produced longer phrases in the happy, interested, and sad contexts. Figure 4 also shows the main effect of repetition. The second (*M*=1198 ms, *SD*=340 ms) and third (*M*=1198 ms, *SD*=329 ms) repetitions were longer than the first (*M*=1153 ms, *SD*=303 ms).

Finally, across both groups, IQ did not significantly correlate with f0 range or mean duration, but there was a significant but weak inverse association with mean intensity ($r=-.09$, $p=.005$). However, because IQ did not differ between the ASD and TD groups, this did not impact any reported group differences.

Experiment 2: Perception of Affective Prosody

Participants

Thirty TD listeners (20 females, 10 males, mean age 22.5 years, age range 18–50 years) and 22 listeners with ASD (2 females, 20 males, mean age 25.9 years, age range 18–43 years) served as participants. The mean IQ of listeners with ASD was 111.3 (IQ range: 88–129), but IQ information was not available for the TD listeners. As in Experiment 1, participants with ASD were recruited through the UTD ARC and had prior DSM IV or DSM-5 diagnoses confirmed via the ADOS by a certified clinician. Five of the listeners with ASD also participated as talkers in Experiment 1. Results in Experiment 2 did not differ between these five participants and the ASD participants who did not participate in Experiment 1. TD participants were undergraduates who received research credit for participating. Each TD listener self-reported no diagnosis of ASD and completed the Broad Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007) to assess their BAP traits. Three listeners had an overall BAPQ score above the BAP cutoff, but their listening test results were included in analysis after individual-level comparisons confirmed that their results closely matched the overall pattern. All listeners were native speakers of American English with no speech or hearing problems. All participants passed a hearing screen at octave frequencies between 0.5 and 4 kHz at ≤ 25 dB HL. ASD participants were financially compensated for their participation, and TD participants were undergraduates who received research credit for being in the study.

Method and Procedure

A subset of the evoked phrases described in Experiment 1 was presented to listeners in an emotion identification task. The listening tests included 330 trials (66 per emotion context); in each trial participants listened to the test phrase over headphones, judged the emotion context by clicking one of the five emotion buttons (“neutral”; “angry”; “happy”; “interested”; and “sad”; displayed in random order per listener), then rated the level of naturalness of the emotional content in the phrase on a slider-type scale ranging from 0 to 10. Identical criteria were used to select test stimuli from the ASD and TD talker groups. For the happy and interested contexts, phrases with the highest f0 range were selected. For the angry context, phrases with the highest median intensity were selected. Sad phrases with the greatest negative f0 slope from the beginning to the end of the phrase were selected. Neutral recordings with the smallest f0 range were selected, corresponding to a relatively flat f0 contour typical of neutral speech. The number of stimuli produced by talkers with ASD and TD in each emotion context was balanced.

Each testing session began with obtaining informed consent, a brief questionnaire and hearing screen, and a practice session. The questionnaire contained questions about the participant’s language and hearing background and history of autism diagnosis (for TD listeners). Diagnoses of ASD for listeners were confirmed by a certified clinician based on the Autism Diagnostic Observation Schedule (Lord, et al., 2000). The practice session consisted of 12 stimuli not included in the main experiment. The task was self-paced, and each session lasted approximately 55 minutes with an optional break at the half-way point. All test procedures were reviewed and approved by the University of Texas at Dallas Institutional Review Board.

Data Analysis

Emotion recognition accuracy (ERA) was calculated as the proportion of emotional phrases correctly identified by listening participants. Mean naturalness ratings were calculated by averaging individual naturalness scores for each emotion category. The data were then summarized for comparisons by test item, talker group and emotion context. ERA and naturalness were treated as dependent variables in separate mixed effects ANOVAs, and talker group, emotion context, phrase and repetition were used as independent variables.

Results

Figure 5 displays ERA by talker group, listener group and emotion context. All mean ERA scores were well above chance level (20%). For ERA, main effects were found for talker group ($F(1,17090)=123.76$, $p<.001$), listener group ($F(1,50)=7.06$, $p<.05$), and emotion

(*F*(4,17090)=127.88, *p*<.001). For the main effect of talker group, emotional phrases produced by talkers with ASD were correctly identified at a significantly higher rate (56%) than TD talkers (48%). For the main effect of listener group, TD listeners had significantly higher ERA (54%) than ASD listeners (49%).

Significant interactions in ERA were found between listener group and emotion context (*F*(4,17090)=9.427, p<.001) and talker group and emotion context (*F*(4,17090)=66.61, p<.001). Bonferroni-corrected post-hoc tests were performed to further examine these interactions. For the interaction between listener group and emotion context, ERA was significantly higher for TD listeners relative to ASD listeners for neutral phrases, and was also higher for the happy context when produced by TD talkers, but TD and ASD listeners did not differ in ERA for any emotion context on phrases produced by ASD talkers. For the interaction between talker group and emotion context, phrases produced by talkers with ASD were more accurately identified than those produced by TD talkers for all emotions except for neutral, with the largest group differences occurring in the angry and sad contexts. For these, ERA was approximately 20 percentage points higher for phrases produced by ASD talkers relative to TD talkers. The reverse pattern, however, was found for neutral phrases – neutral phrases produced by TD talkers were over 10% more likely to be identified compared to those produced by talkers with ASD.

Figure 6 shows mean listener naturalness ratings for each talker group and emotion context. For naturalness ratings, main effects of talker group (*F*(1,17090)=98.91, *p*<.001) and emotion context (*F*(4,17090)=141.60, *p*<.001) were found. Phrases produced by talkers with ASD received lower mean naturalness ratings (5.74) than those produced by TD talkers (6.05). Angry phrases were rated as the most natural sounding (6.31) and neutral phrases were rated as the least natural (5.10).

Interactions in mean naturalness scores were found between talker group and listener group (*F*(1,17090)=10.79, *p*<.01), which was driven by ASD listeners rating TD talkers (M=6.35, *SD*=2.18) as more natural than ASD talkers (M=5.91, *SD*=2.60) to a greater degree than TD listeners did (M=5.84, *SD*=2.27 for TD talkers compared to M=5.62, *SD*=2.49 for ASD talkers). Interactions were also found between listener group and emotion context (*F*(4,17090)=34.62, *p*<.001), and talker group and emotion context (*F*(4,17090)=11.14, *p*<.001). When the naturalness ratings were broken down by emotion context, happy and interested phrases produced by TD talkers were rated more natural by both groups of listeners, compared to those produced by ASD talkers. In addition, there were significant listener group differences for neutral, interested and sad phrases. A large listener group effect was found for neutral phrases – those heard by TD listeners were rated much more natural sounding compared to those heard by listeners with ASD. Sad and interested phrases produced by TD talkers were rated more as sounding more natural by listeners with ASD than TD listeners. No talker group or listener group differences in naturalness were found for angry phrases. IQ was not available for TD listeners, but ERA was positively correlated with IQ (r = .08, *p* < .001) and negatively correlated with naturalness ratings (r = −0.28, *p* = <.001) for ASD listeners. ERA did not differ between male and female listeners, but females (*M*=7.74, *SD*=2.53) did rate phrases as more natural than did males (*M*=5.97, *SD*=2.34;

F(1,7258)=339.88, *p*<.001). The small number of females in the ASD group did not allow examination of whether these patterns differed between the two groups.

General Discussion

This study examined the production and perception of affective prosody in adult males with ASD and TD controls. Compared to controls, participants with ASD produced emotional phrases that had increased f0 range (i.e., more pitch variability), greater mean intensity (i.e., louder), and longer duration. However, when these emotional phrases produced by ASD and TD talkers were then presented to ASD and TD listeners to determine whether affective prosodic differences in ASD affects emotion recognition accuracy on the part of potential social partners, phrases produced by talkers with ASD were *more* accurately identified than those produced by TD talkers. Thus, although atypical in f0 variability, intensity, and duration, the affective prosody produced by the adults with ASD nevertheless contained identifiable acoustic cues of emotion that aided in their recognition. This likely occurred because the prosodic differences for the ASD group were quantitatively, not qualitatively, different from TD speech, and their exaggerated presentations of typical emotional acoustic cues may have facilitated emotion identification on the part of listeners. These patterns occurred across all emotional categories except for the neutral context, where prosody did not differ between the two groups, suggesting that the prosodic differences reported for ASD were specific to the expression of emotion and did not occur for phrasing lacking affective expression.

Importantly, the finding of increased f0 range and greater intensity for emotional speech produced by the ASD group is inconsistent with traditional conceptualizations of voices in ASD being characterized as “flat” or “monotone”. Rather, they join a growing number of studies that challenge the assumption of ASD as being uniformly characterized by diminished emotional expressivity, including reports indicating a higher fundamental frequency and wider pitch range in laughter (Hudenko and Magenheimer, 2011) and expressive speech (Fosnot and Jun 1999; Nadig and Shaw, 2012), as well as recent findings of exaggerated facial affect in ASD (Faso, Sasson & Pinkham, 2015). The exaggerated emotional speech production in ASD reported here accurately conveyed emotional information to listeners but nevertheless led to judgments of being perceived as less natural compared to speech produced by TD talkers. In this way, the findings reported here suggest differences, but not a reduction, in affective speech production in ASD that could potentially impact social evaluation and social interaction quality despite not impairing transmission of categorical emotional information.

Although the current study cannot address the origin of the atypical emotional prosodic patterns reported here for ASD talkers, the findings are consistent with a general speech attunement difference where individuals with ASD have difficulty emulating socially-normative standards due to difficulties with social understanding, reciprocity, and communication intent (Shriberg et al., 2011). From this perspective, individuals with ASD may be less adept at detecting discrepancies between their own and more typical speech styles, and therefore do not attune their accordingly (Paul et al., 2008; Shriberg et al., 2011; Diehl & Paul, 2012). Alternatively, they may detect these differences but lack the motivation or interest to do so. Although the current study was not designed to differentiate between these two explanations, ASD listeners, like TD listeners, reliably rated the speech of ASD talkers as less natural than the speech of TD talkers. This suggests that individuals with ASD are sensitive to acoustic differences in emotional speech and make evaluative judgments about them similarly to TD controls. Thus, the evidence provided here suggests that their divergent patterns of speech production in f0, intensity and duration are likely not driven by a lack of ability to detect these cues.

Indeed, although listeners with ASD exhibited decreased ability to identify emotions in speech, a finding consistent with previous studies (Paul et al., 2005; Peppé et al., 2007; Stewart et al., 2012), this effect was small, and was driven by group differences in the neutral context and for happy phrases produced by TD talkers. Overall, the group differences in emotional speech perception were far less robust than the differences in emotional speech production. This interpretation is consistent with emerging evidence that individuals with ASD may differ from TD controls to a greater degree in the expression of emotion and non-verbal cues than in the perception of them (Grossman and Tager-Flusberg, 2012).

The current study has several limitations. First, listening test stimuli were only produced by male talkers. Female talkers were excluded due to an inability to recruit female participants with ASD in large enough numbers, and because sex-related anatomical and social factors can lead to differences in f0 ranges and use of affective prosody. Future studies are encouraged to determine whether the findings reported here extend to females with ASD. Second, there were more female listeners in the TD group than the ASD group in Experiment 2. Although emotion recognition accuracy did not differ by listener gender across the groups, the small number of females in the ASD group precluded examination of whether effects might differ between males and female listeners with ASD. Third, expressive speech was produced in a laboratory setting to maximize control and standardize recording procedures, but such speech may differ from expressive speech in everyday settings. However, our use of a naturalistic evocation procedure, coupled with ERA rates well above chance for all emotional categories, suggest that the emotional speech produced by participants was a valid representation of emotional expressivity. Finally, the ASD participants included here were adults with intellectual ability in the normal range, and it is unclear whether our findings would extend to children or lower functioning individuals. It may be the case that prior descriptions of flat affect in ASD are more characteristic of a subtype of ASD not represented in the current study. Future research addressing this question is warranted. Further, although IQ did not differ between ASD and TD talkers in Experiment 1, IQ information was not available for TD listeners in Experiment 2. However, because TD participants in Experiment 2 were drawn from the same university subject pool as in Experiment 1, it is unlikely that their IQ differed notably from those in Experiment 1.

Despite these limitations, the current study advances understanding of production and perception of affective speech in ASD. The elicitation procedure used here resulted in ecologically valid emotional speech production that reliably differed in ASD relative to TD controls. Emotional speech in ASD was characterized by greater pitch variability, increased intensity, and longer durations. These differences were detected by both TD and ASD listeners, who rated emotional speech in ASD as less natural but nevertheless identified the emotion in ASD speech with greater accuracy. How these differences in emotional speech production develop, and whether they affect social interaction and relate to social functioning more broadly, are open questions worthy of investigation.

Acknowledgments

The authors wish to thank the individuals who participated in the study, and the Nonpareil Institute of Plano, TX for helping with participant recruitment. We thank the reviewers for their comments on a previous version of the manuscript. This research was funded in part by a grant from the National Institute of Mental Health at National Institutes of Health (R15 MH1015945, PI Sasson). This work is based on a doctoral dissertation by DH, who was supported in part by a grant from the National Science Foundation (1124479, PI Assmann).

References

Banse R, & Scherer K (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70, 614–636.8851745

Bänziger T, Mortillaro M, & Scherer K (2012). Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. Emotion, 12, 1161–1179.22081890

Bates D, Maechler M, Bolker B, Walker S (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. doi:10.18637/jss.v067.i01.

Begeer S, Koot HM, Rieffe C, Terwogt M, & Stegge H (2008). Emotional competence in children with autism: Diagnostic criteria and empirical evidence. Developmental Review, 28(3), 342–369.

Boersma P, & Weenink D (2014). Praat: doing phonetics by computer [Computer program]. *Version 5.4.01.* Retrieved December 30, 2014, from http://www.praat.org/

Diehl JJ, & Paul R (2012). Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders. Research in Autism Spectrum Disorders, 6, 123–134.22125576

Fairbanks G & Provonost W (1939). An experimental study of the pitch characteristics of the voice during the expression of emotions. Speech Monographs, 6, 87–104.

Faso DJ, Sasson NJ, & Pinkham AE (2015). Evaluating posed and evoked facial expressions of emotion from adults with autism spectrum disorder. Journal of Autism and Developmental Disorders, 45, 75–89.25037584

Fosnot SM, & Jun SA (1999). Prosodic characteristics in children with stuttering or autism during reading and imitation. Paper presented at the 14th annual congress of phonetic sciences, 1925–1928.

Golan O, Baron-Cohen S, Hill JJ, & Rutherford MD (2007). The 'Reading the Mind in the Voice' Test-Revised: A study of complex emotion recognition in adults with and without autism spectrum conditions. Journal of Autism and Developmental Disorders, 37, 1096–1106.17072749

Green H, & Tobin Y (2009). Prosodic analysis is difficult but worth it: A study in high functioning autism. International Journal of Speech-Language Pathology, 11, 308–315.

Grossman RB, & Tager-Flusberg H (2012). Quality matters! Differences between expressive and receptive non-verbal communication skills in adolescents with ASD. Research in Autism Spectrum Disorders, 6 (3), 1150–1155.22773928

Hubbard K, & Trauner DA (2007). Intonation and emotion in autism spectrum disorders. Journal of Psycholinguistic Research, 36, 159–173.17136465

Hudenko WJ, & Magenheimer MA (2011). Listeners prefer the laughs of children with autism to those of typically developing children. Autism, 16 (6), 641–665.21810911

Hurley R, Losh M, Reznick J, & Piven J (2007). The Broad Autism Phenotype Questionnaire. Journal of Autism and Developmental Disorders, 1679–1690.17146701

Hus V, & Lord C (2014). The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores. Journal of Autism and Developmental Disorders, 44 (8), 1996–2012.24590409

Juslin PN, & Laukka P (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin, 129, 770–814.12956543

Kawahara H, Masuda-Katsuse I, & de Cheveigné A (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction. Speech Communication, 27, 187–207.

Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, Pickles A, Rutter M (2000). The Autism Diagnostic Observation Schedule–Generic: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of Autism and Developmental Disorders, 30, 205–223.11055457

Lyons M, Simmons ES, & Paul R (2014). Prosodic development in middle childhood and adolescence in high-functioning autism. Autism Research, 7, 181–196.24634421

Murray IR, & Arnott JL (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustical Society of America, 93, 1097–1108.8445120

Nadig A, & Shaw H (2012). Acoustic and perceptual measurement of expressive prosody in high-functioning autism: Increased pitch range and what it means for listeners. Journal of Autism and Developmental Disorders, 499–511.21528425

Paul R, Augustyn A, Klin A, & Volkmar FR (2005). Perception and production of prosody by speakers with autism spectrum disorders. Journal of Autism and Developmental Disorders, 35, 205–220.15909407

Paul R, Bianchi N, Augustyn A, Klin A, & Volkmar FR (2008). Production of syllable stress in speakers with autism spectrum disorders. Research in Autism Spectrum Disorders, 2, 110–124.19337577

Peppé S, & McCann J (2003). Assessing intonation and prosody in children with atypical language development: The PEPS-C test and the revised version. Clinical Linguistics and Phonetics, 17, 345–354.12945610

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

R Core Team. (2015). R: A language and environment for statistical computing. Retrieved from R Foundation for Statistical Computing: http://www.R-project.org

Russell JA (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161–1178.

Russell JA (2003). Core affect and the psychological construction of emotion. Psychological Review, 110, 145–172.12529060

Rutherford MD, Baron-Cohen S, & Wheelwright S (2002). Reading the mind in the voice: A study with normal adults and adults with Asperger Syndrome and high functioning autism. Journal of Autism and Developmental Disorders, 32, 189–194.12108620

Sasson NJ, Faso DJ, Nugent J, Lovell S, Kennedy DP, & Grossman RB (2017). Neurotypical peers are less willing to interact with those with autism based on thin slice judgments. Scientific Reports, 7: 40700. doi:10.1038/srep40700.28145411

Sasson NJ, Lam KS, Childress D, Parlier M, Daniels JL, & Piven J (2013). The Broad Autism Phenotype Questionnaire: Prevalence and Diagnostic Classification. Autism Research, 134–143.23427091

Sasson NJ, Pinkham AE, Carpenter KH, & Belger A (2011).The benefit of directly comparing autism and schizophrenia for revealing mechanisms of social cognitive impairment. Journal of Neurodevelopmental Disorders, 3(2), 87–100. doi:10.1007/s11689-010-9068-x.21484194

Scherer KR (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In Izard CE, Emotions in Personality and Psychpathology (pp. 495–529). New York: Plenum Press.

Scherer KR (1986). Vocal affect expression: A review and a model for future research. Psychological Bulletin, 2, 143–165.

Scherer KR (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication, 40, 227–256.

Shriberg LD, Paul R, Black L, & Van Santen J (2011). The hypothesis of apraxia of speech in children with autism spectrum disorder. Journal of Autism and Developmental Disorders, 41, 405–426.20972615

Stewart ME, McAdam C, Ota M, Peppé S, Cleland J (2012). Emotional recognition in autism spectrum conditions from voices and faces. Autism, 17, 6–1423045218

Uljarevic M, & Hamilton A (2013). Recognition of emotions in autism: a formal meta-analysis. Journal of Autism and Developmental Disorders, 43(7), 1517–1526.23114566

Van Bourgondien ME, & Woods A (1992). Vocational possibilities for high-functioning adults with autism In Schopler E, & Meisbov G (Eds.), High-functioning individuals with autism (pp. 227–242). New York: Plenum Press.

Wang AT, Lee SS, Sigman M, & Dapretto M (2007). Reading affect in the face and voice: Neural correlates of interpreting communicative intent in children and adolescents with autism spectrum disorders. Archives of General Psychiatry, 64, 698–708.17548751

Wechsler D (1999). Wechsler abbreviated scales of intelligence (WASI). San Antonio, TX: The Psychological Corporation/Harcourt Assessment.

Williams CE & Stevens KN (1972). Emotions and speech: Some acoustical correlates. Journal of the Acoustical Society of America, 52(4), 1238–1250.4638039

Wundt W (1909). Grundriss der psychologie, Achte auflage (Outlines of psychology). Leipzig, Germany: Engelmann.

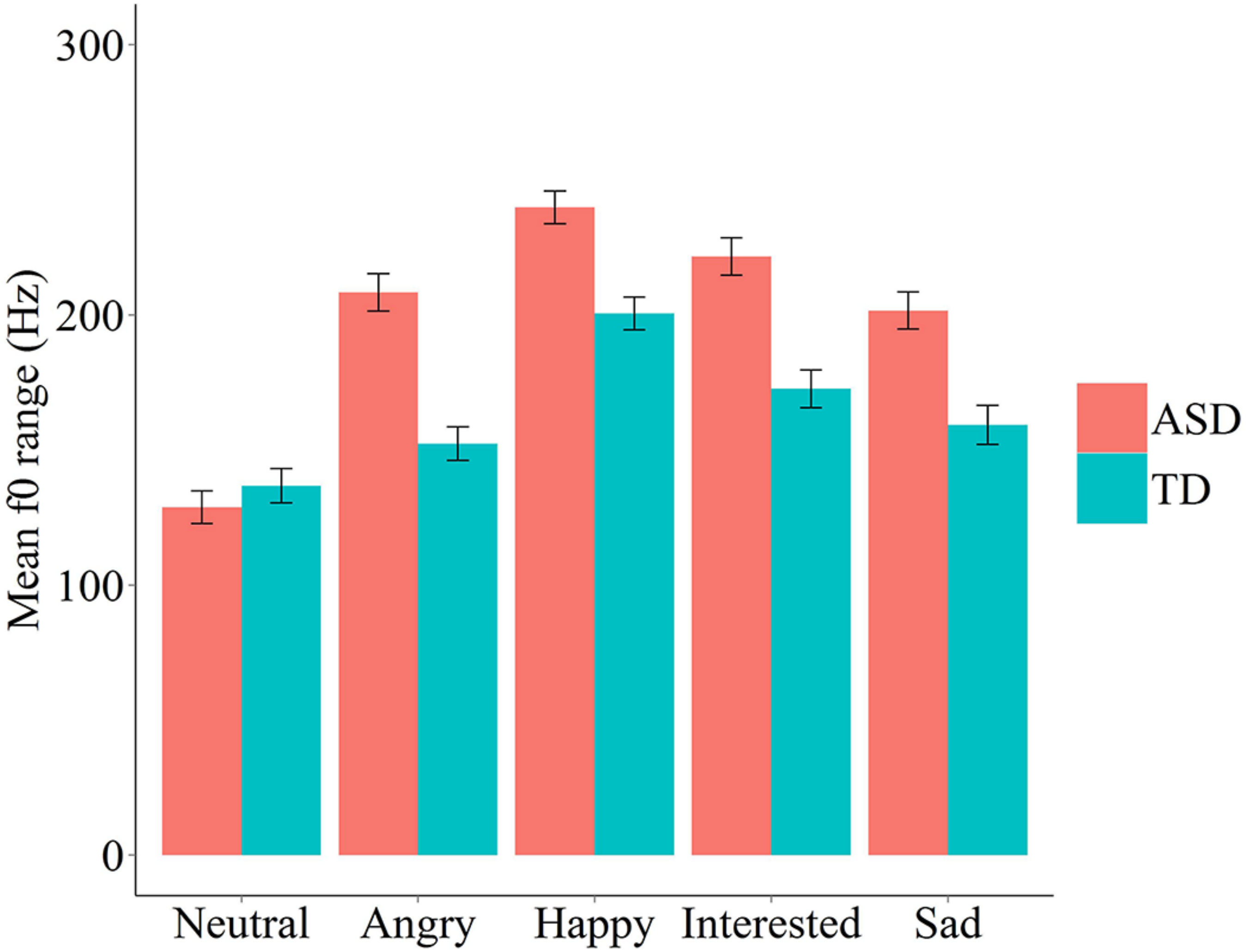

Figure 1.

Mean f0 range (in Hz) for evoked emotional phrases produced by talkers with ASD and TD controls. Bars represent talker group f0 range means in each emotion context (collapsed across repetition). Error bars represent +/− one mean standard error.

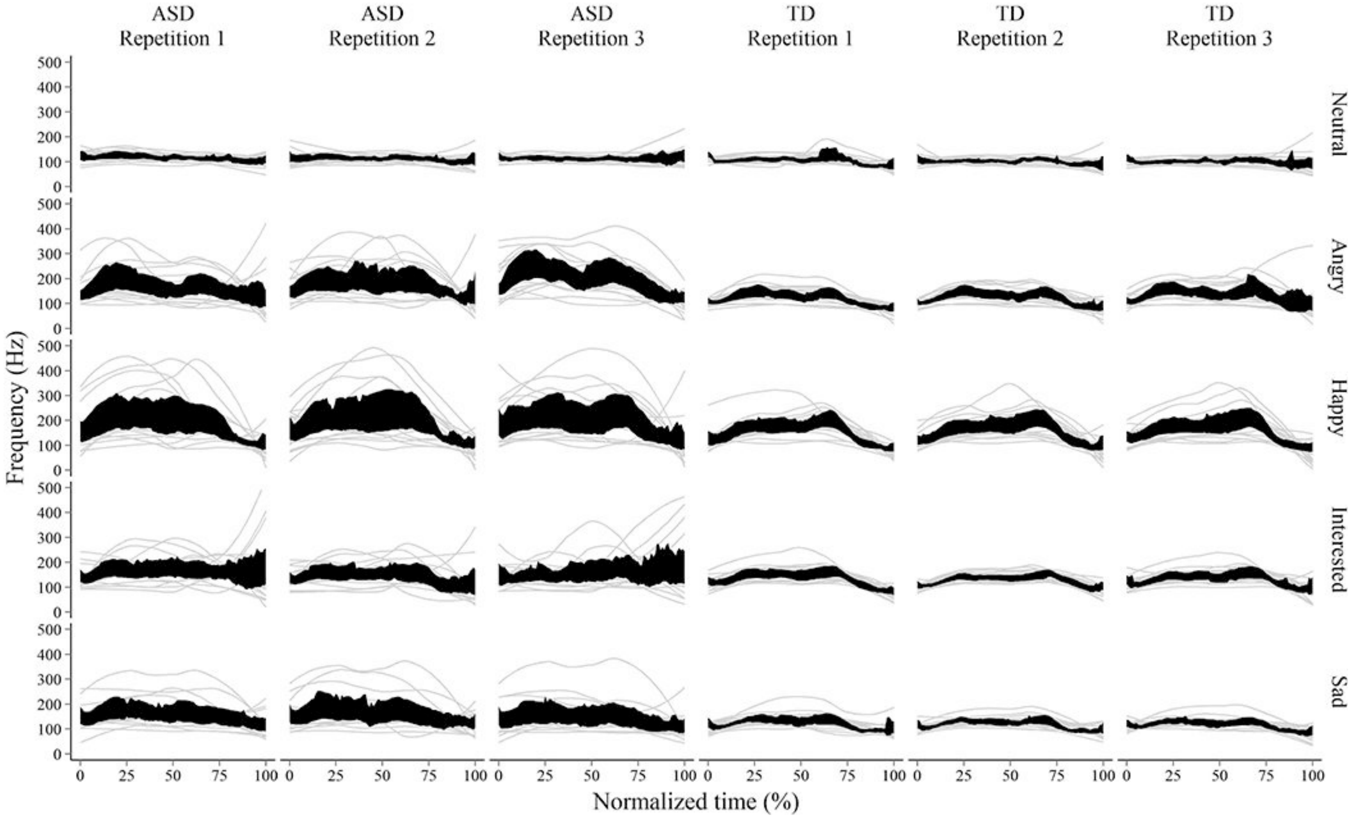

Figure 2.

F0 contours for individual evoked phrases (light grey) and 95% confidence bands (superimposed in black) by talker group for each emotion context and repetition for the phrase “Why did you do that?” The duration of each phrase was time-normalized so that the beginning and end of each phrase were aligned.

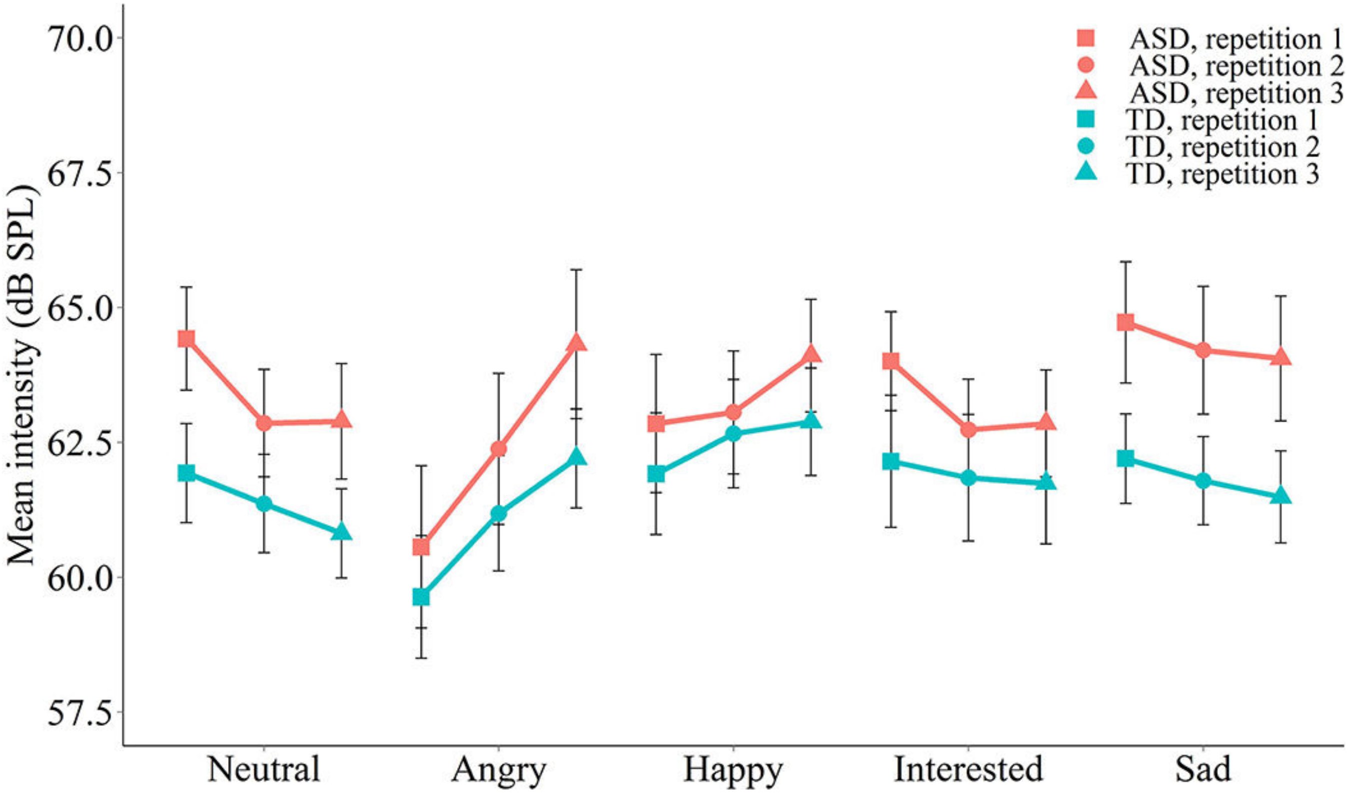

Figure 3.

Mean intensity levels (dB SPL) for evoked phrases produced by talkers with ASD and TD controls. Points represent means for each talker group, emotion context and repetition. Error bars represent +/− one mean standard error.

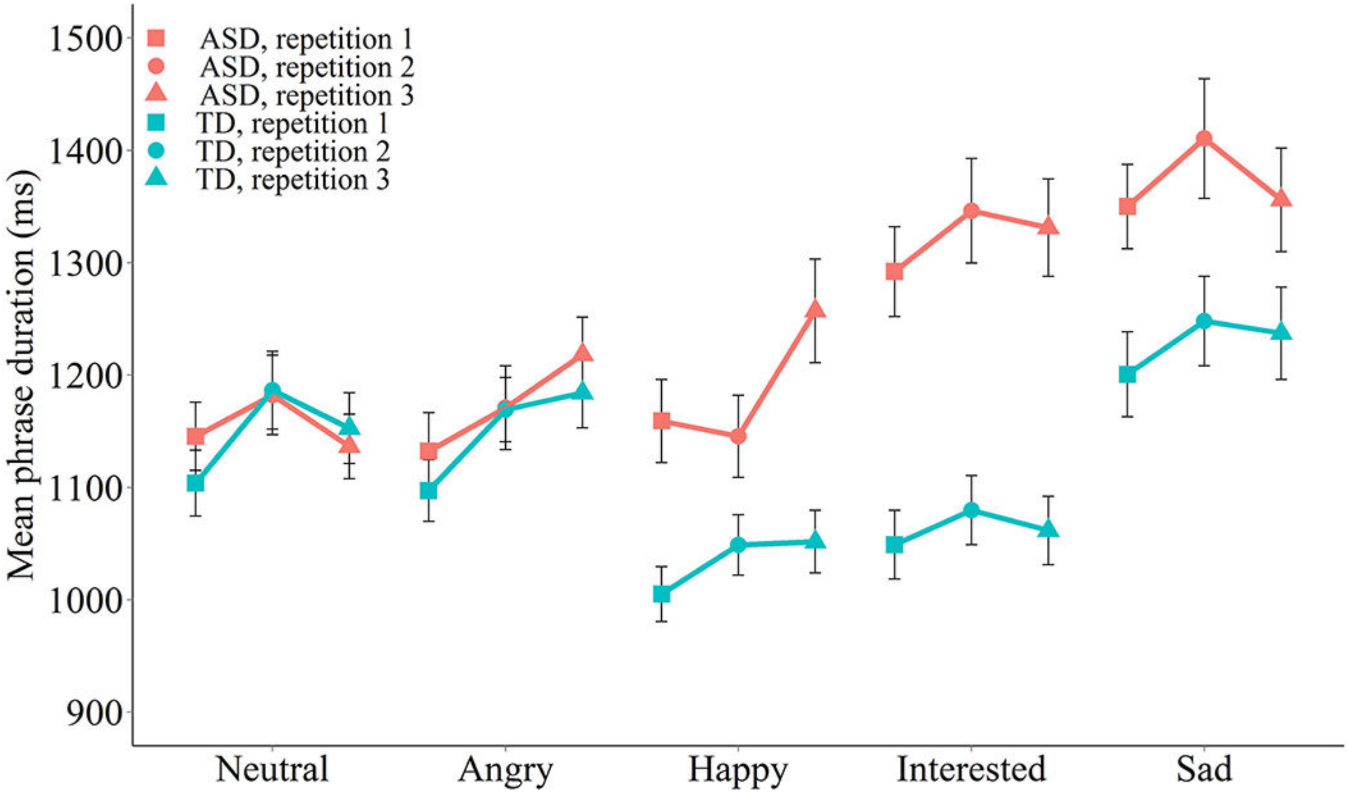

Figure 4.

Mean phrase duration (in ms) for evoked phrases produced by talkers with ASD and TD controls. Points represent means for each talker group, emotion context and repetition. Error bars represent +/− one mean standard error.

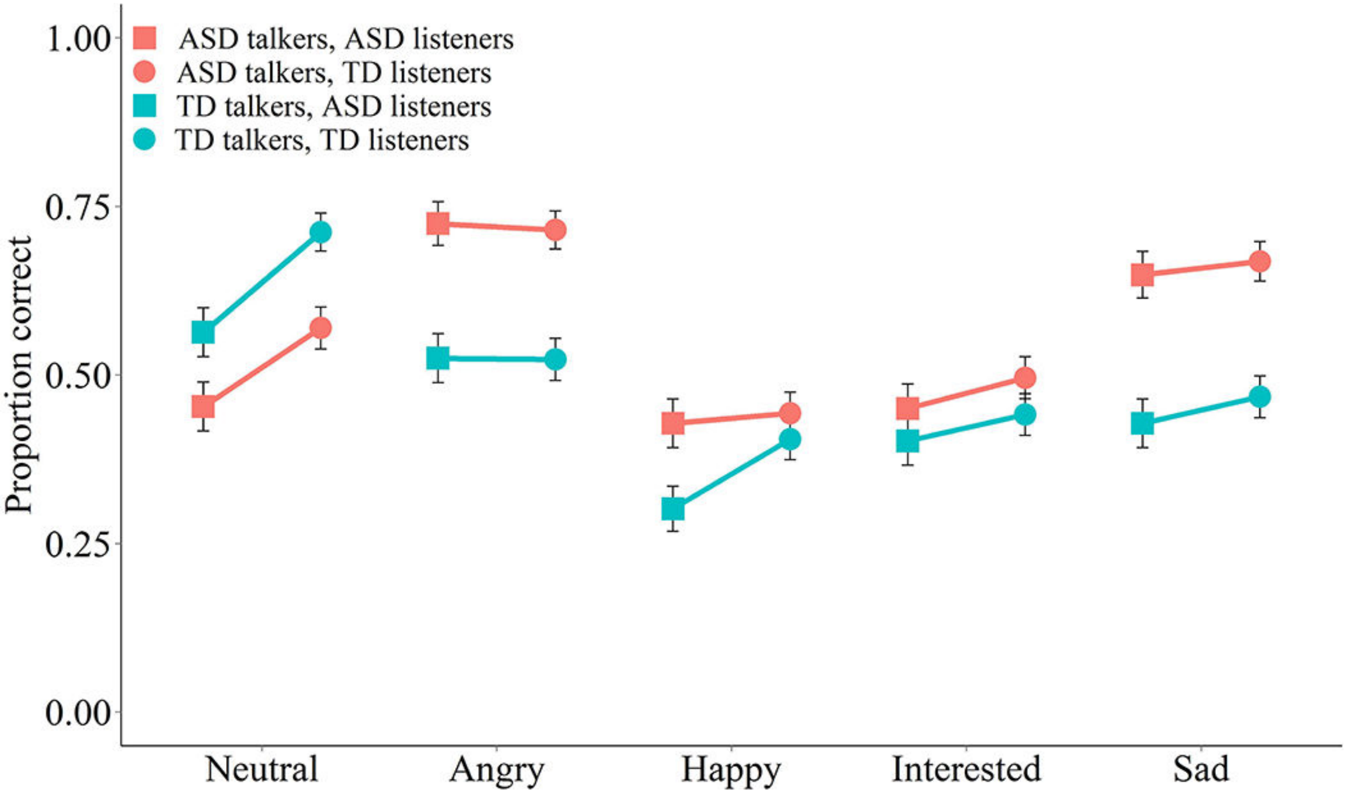

Figure 5.

Mean listener emotion recognition accuracy (ERA) scores for evoked phrases produced by talkers with ASD and TD controls. Points represent mean scores for each emotion context, talker group, and listener group. Error bars represent +/− one 95% confidence interval around the mean.

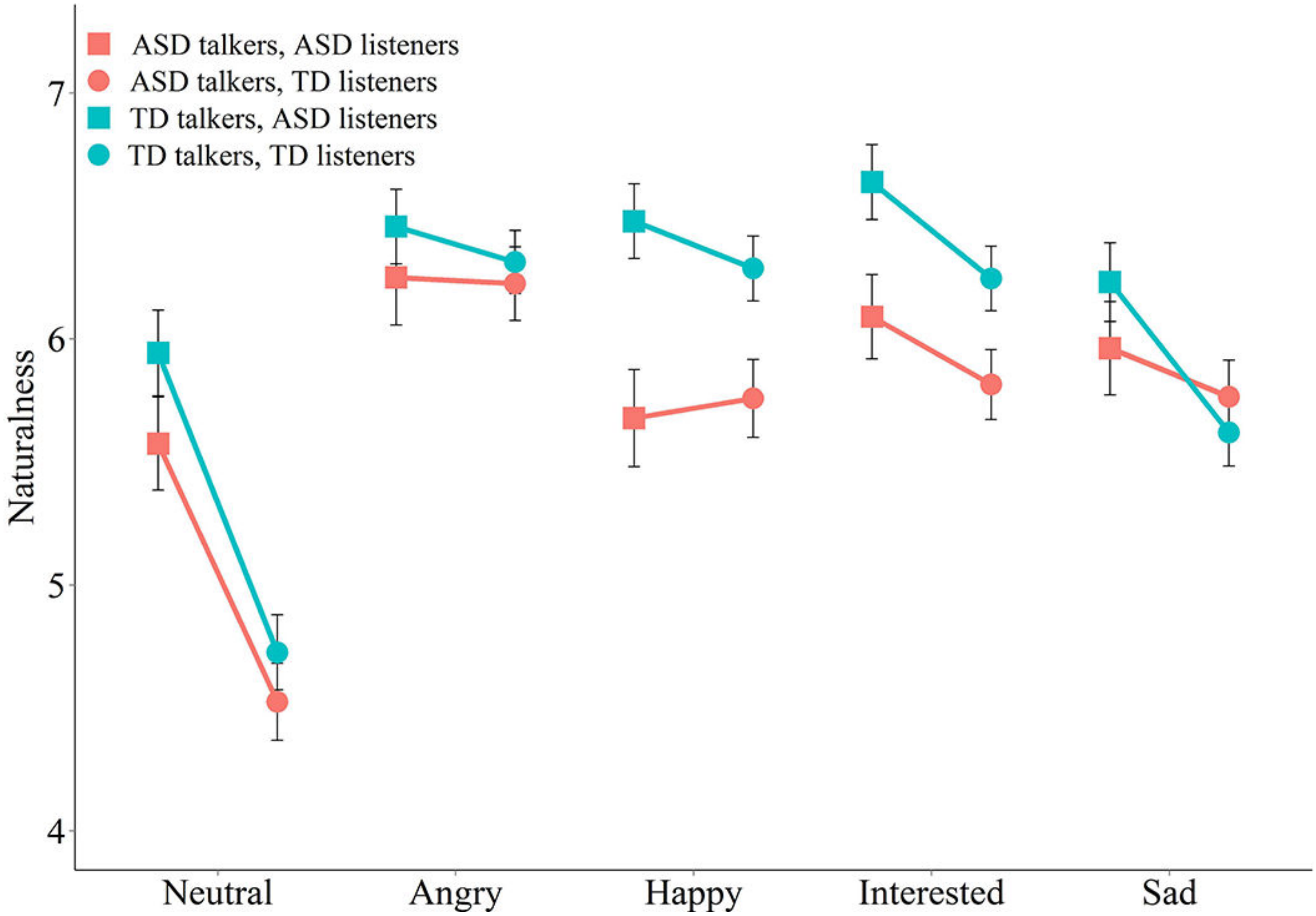

Figure 6.

Mean listener naturalness ratings for evoked phrases produced by talkers with ASD and TD controls. Points represent mean scores for each emotion context, talker group, and listener group. Error bars represent +/− one 95% confidence interval around the mean.

Table 1.

Evoked phrase acoustic data. Group means, standard deviations, and the mean range of the 225 phrases produced in each emotion context by each talker group for f0, intensity, and duration.

		f0 (Hz)			Intensity (dB SPL)			Duration (ms)		
context	**group**	**mean**	**SD**	**range**	**mean**	**SD**	**range**	**mean**	**SD**	**range**
neutral	ASD	119.0	46.4	54.2	65.4	3.7	20.0	1154.7	273.6	1943.0
neutral	TD	108.3	61.7	66.0	63.6	3.0	17.5	1147.7	277.1	1472.0
angry	ASD	176.4	86.9	141.0	67.3	4.2	26.7	1173.8	303.6	1618.0
angry	TD	128.0	56.4	91.0	66.0	3.3	19.2	1150.1	254.0	1319.0
happy	ASD	198.8	106.7	166.4	67.5	3.4	19.8	1187.2	349.3	2360.0
happy	TD	159.4	74.8	129.3	66.4	3.4	17.3	1035.1	228.9	1174.0
interested	ASD	158.0	89.3	130.1	65.4	3.5	19.8	1323.0	372.1	1931.0
interested	TD	134.7	62.2	97.9	65.4	4.2	19.0	1063.4	264.3	1427.0
sad	ASD	150.6	73.9	104.8	65.7	4.4	21.4	1372.2	398.2	3491.0
sad	TD	122.2	61.8	79.5	63.2	3.3	16.6	1228.6	341.9	1578.0