Emotional recognition in autism spectrum conditions from voices and faces

Mary E. Stewart
Applied Psychology, Heriot-Watt University, Edinburgh, Scotland, EH14 4AS, UK

Clair McAdam
Applied Psychology, Heriot-Watt University, Edinburgh, Scotland, EH14 4AS, UK

Mitsuhiko Ota
Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, Scotland, EH8 9AD, UK

Sue Peppé
Speech and Hearing Sciences, Queen Margaret University, Edinburgh, Scotland, EH21 6UU, UK

Joanne Cleland
Speech and Hearing Sciences, Queen Margaret University, Edinburgh, Scotland, EH21 6UU, UK

Abstract
The present study reports on a new vocal emotion recognition task and assesses whether people with autism spectrum conditions (ASC) perform differently from typically developed individuals on tests of emotional identification from both the face and the voice. The new test of vocal emotion contained trials in which the vocal emotion of the sentence were congruent, incongruent, or neutral with respect to the semantic content. We also included a condition in which there was no semantic content (an 'mmm' was uttered using an emotional tone). Performance was compared between 11 adults with ASC and 14 typically developed adults. Identification of emotion from sentences in which the vocal emotion and the meaning of sentence were congruent was similar in people with ASC and a typically developed comparison group. However, the comparison group was more accurate at identifying the emotion in the voice from incongruent and neutral trials, and also from trials with no semantic content. The results of the vocal emotion task were correlated with performance on a face emotion recognition task. In decoding emotion from spoken utterances, individuals with ASC relied more on verbal semantics than did typically developed individuals, presumably as a strategy to compensate for their difficulties in using prosodic cues to recognize emotions.

Keywords
Emotion, autism spectrum conditions, prosody, vocal emotion
Individuals with autism spectrum conditions (ASC) show divergent behaviours in processing and understanding emotional states in others. This has been demonstrated in emotion identification either from the face or the voice (Balconi and Carrera, 2007; Boucher et al., 1998; Hobson et al., 1989; Humphreys et al., 2007; Loveland et al., 1995; Peppé et al., 2007; Van Lancker et al., 1989). Problems in identifying vocal emotion have been reported in both children and adults with ASC. For instance, children with autism have difficulty matching vocally expressed affect to static facial expressions or to emotion words (Boucher et al., 1998; Hobson et al., 1989; Van Lancker et al., 1989); and vocal emotions to dynamic images such as videos of faces (Loveland et al., 1995). Peppé and colleagues found that children with high-functioning autism (HFA) were less accurate at identifying emotion in a voice (Peppé et al., 2007). In their task, the children heard a word such as ‘salad’ said either in a happy or sad voice, which they had to match with a happy or a sad face shown on a computer screen. Paul and colleagues similarly found identification of affect from the voice to be impaired in children with ASC (Paul et al., 2005). In their study the children heard a sentence and were asked to indicate a picture that represented the emotion being portrayed in the voice. For instance, the children might hear ‘You’re going to be late for school’ said in a calm or an excited voice, and then see a picture of a lady looking calm and a picture of a lady feeling excited to choose from.

Similar differences have been found in adults with HFA/Asperger Syndrome (AS) (Golan et al., 2006; Golan et al., 2007; Rutherford et al., 2002). For example, Rutherford and colleagues (Rutherford et al., 2002) played recorded segments of dialogue from dramatic audio books (e.g. ‘No, honestly I do’), from which participants were asked to identify the speaker’s mental attitude or emotion from a choice of two (e.g. ‘Earnest’ vs. ‘Alarmed’). The ASC group were less accurate at identifying the emotion in the voice than comparison groups.

Such differences are found in emotion identification based not only on vocal information but also on facial expressions. Golan et al. (2006) used the Cambridge mindreading face-voice battery, in which participants were played either a short neutral sentence spoken by an actor in a particular emotional tone, or a short clip of actors portraying a facial emotion. Participants were instructed to choose one response from four that best described the speaker’s mental state. The results showed that, compared with typically developed general population adults, adults with Asperger Syndrome (AS) were less accurate at identifying emotion from both voices and faces. Kleinman and colleagues showed a similar result in high functioning adults with ASC using the Mental State Voices Task (Kleinman et al., 2001). A neutral sentence was said in one of the six ‘basic’ emotions (happy, sad, angry, afraid, surprised and disgusted) (Ekman, 1993) and in six more complex emotions (arrogant, guilty, calm, anxious, bored and interested). Participants were also given the Reading the Mind in the Eyes task (Baron-Cohen et al., 2001).

Adults with ASC were impaired compared with typically developed individuals on both tasks. Other studies also show that both adults and children with ASC tend to be less accurate at identifying emotions from faces (Balconi and Carrera, 2007; Humphreys et al., 2007). The difficulty shown by individuals with ASC increases as the emotion to be detected becomes more complex than the ‘basic’ categories of happiness, sadness, fear, anger, surprise and disgust (Capps et al., 1992).

However, individuals with ASC do not consistently diverge from neurotypicals in their decoding of emotions. In Grossman et al. (2000), children and adolescents with AS did not differ from comparison groups in their ability to recognize facial emotions when the faces were paired with matching words (e.g. a happy face with the word HAPPY), even though they did so when they were shown mismatching words (e.g. a happy face with the word AFRAID), which typically led them to a response based on the word rather than the face. Similarly, Lindner and Rosén (2006) found that children and adolescents with AS did not perform worse than their typically developing
counterparts in deciding the emotion expressed in an acted out scene when the emotion was expressed in the semantic content of the actor’s verbalization. When this verbal semantic cue was absent, the AS participants encountered difficulties in identifying the intended emotion. These findings indicate that individuals with ASC rely on linguistic semantics, where their emotion decoding ability is spared.

The main purpose of this study was to test whether a similar compensatory strategy could be observed in emotion detection in speech stimuli only (without any accompanying facial expression or visual context). We expect individuals with ASC to show more variation than typically developed individuals in their emotional interpretation of auditory speech samples depending on how much information on emotion is directly encoded in the linguistic material (e.g. words that denote a particular emotional state). To test this, we developed measures of emotional interpretation using sentences with different degrees of congruency between prosodic and semantic information and also hummed utterances (‘mmm’) devoid of semantic content. Each sentence or ‘mmm’ was said in a different emotion: happy, anger, fear, surprise or disgust. The intended emotion in the voice was in some trials also discernable from the semantic content of the sentence (e.g. ‘Stop that at once!’ said in an angry voice; the voice and the meaning of the sentence were congruent), but in contradiction to the semantic implications in others (e.g. ‘Stop that at once!’ said in a happy voice; the voice and the meaning of the sentence were incongruent). We predicted a difference between individuals with ASC and typically developed participants in the latter (‘incongruent’) condition, but a reduced effect or no difference in the former (‘congruent’) condition, where the meaning of the sentence can be used to identify the emotion expressed. We also predicted a difference between the two groups in the ‘mmm’ condition, as the only source of emotional information in these materials was the prosody. In this condition, individuals with ASC would not be able to use a semantic strategy to interpret the emotional intonation.

A second goal of the study was to examine the connection between vocal and facial recognition of emotion. If the emotional processing difficulties in ASC are caused by a deficit in similar cognitive processes, then emotion interpretations of different sources, such as faces and vocalization, will be subjected to the same underlying problem. We would therefore expect to see a close connection between these behaviours (understanding of emotions in faces and in vocalization) in each ASC individual, as has been found in previous studies (Golan et al., 2006; Kleinman et al., 2001). To test this, we also administered a task in which participants were required to identify emotion from a face. The prediction was that for individuals with ASC, there would be a correlation between scores in the facial emotion recognition task and the incongruent condition of the vocal task.

Method

Participants

The participants were 11 individuals with ASC and 14 typically developed young adults between 17 and 39 years old. Participants with ASC were recruited from a ‘one-stop shop’ facility in Edinburgh for individuals with AS and HFA. The facility provides free social opportunities, one-to-one advice and support in a range of areas, including outreach and housing. Only those participants who had been previously diagnosed as meeting the criteria of Diagnostic and Statistical Manual of Mental Disorders Edition IV (DSM-IV) for autism spectrum disorders by a qualified clinician were included in the study. Only those participants who had notes outlining their diagnosis were included in the study. Comparison participants were recruited from the general population.
**Measures**

**Tests of emotional prosody.** The participants listened to recorded stimuli played on a computer over headphones through the platform of E-Prime. The stimuli were recorded by a professional female actor. The actor had a standard Scottish accent and was approximately the same age as the participants. There were three tests of prosody assessing the impact of semantic/lexical information and prosodic factors. The stimuli were said in one of five emotions: anger; fear; happiness; surprise; and disgust. Participants were asked to ‘...indicate the emotion that is being portrayed in the voice’ in each stimulus by selecting one of six numbered answer options that appeared on the screen: Anger; Fear; Happy; Surprise; Sad; Disgust. Participants selected the correct answer using the corresponding number on the keyboard. The answer options were randomized to ensure that reaction time was not influenced by the number position of the correct answer.

**Sentences.** Participants listened to 120 sentences the meaning of which was either emotional or neutral. Emotional sentences were sentences that implied an emotional state; for example, ‘Stop that at once!’ implies an angry emotional state. Sentences were tested in a pilot study where participants were asked to identify the emotional state in the sentence. Sentences in which the emotion was identified at less than 75% accuracy were discarded. Eight sentences were chosen for each of the 5 emotions being tested (happiness, fear, surprise, anger and disgust). These sentences were said with either emotionally congruent or emotionally incongruent prosody. For example, ‘Stop that at once!’ said with prosody implying anger would be congruent, but if it was said with prosody implying happiness it would be incongruent. Neutral sentences that did not imply any emotional state were also included, for example, ‘He drank a cup of tea’. Eight neutral sentences were chosen and these eight sentences were each heard portraying each of the five emotions. In total, each participant heard 40 congruent, 40 incongruent and 40 neutral sentences.

‘mmm’s. To test the ability to identify vocally expressed emotion without lexical information, participants listened to vocalizations of ‘mmm’ uttered in each of the five emotions. There were a total of 25 trials, and the answer selection process was the same as that for the sentences and words. Participants were given five practice trials before the real test.

**Facial emotion recognition.** For the facial emotion recognition task, eight faces were chosen for each of the six basic emotions from the JAFFE database (Lyons et al., 1999). Participants viewed the faces on the computer screen and were asked to choose which emotion the face was portraying from a list of the six basic emotions that appeared alongside the picture.

**Procedure**

Participants were asked to complete the tests in a set order. Participants completed the Mill Hill vocabulary test first, followed by tests of emotional prosody and the facial emotion recognition test. Participants could take regular breaks, all the testing was completed in one session.

**Results**

The groups were matched on verbal IQ ($t = 1.547$, $df = 23$, n.s.), which was measured using the multiple choice form of the Mill Hill Vocabulary Scale (Raven et al., 1998), and on age. Gender ratios in the two groups were similar. Four of the group with ASC were
female and seven were male, six of the typically developed group were female and eight were male.

Accuracy and reaction time for the voice tasks, mean performance on the faces, age and verbal ability tasks are shown in Table 1. In the voice task, accuracy was much higher for both groups when the emotion in the voice was congruent with the meaning of the sentence. A repeated measures analysis of variance was carried out to determine the effects of group and condition in the sentences task. There was a main effect of congruency ($F_{2,22} = 84.71, p < 0.001$), a main effect of group ($F_{1,23} = 29.30, p < 0.001$), and an interaction between group and congruency ($F_{2,22} = 4.43, p < 0.05$). Where data did not meet parametric assumptions Mann-Whitney U tests were completed on between group comparisons. There were no significant differences between the groups in the congruent condition. However, the comparison group was more accurate than the ASC group in the neutral ($Z = 3.90, n = 25, p < 0.001$) and incongruent conditions ($Z = 3.77, n = 25, p < 0.001$) (Figure 1 and Table 1). The comparison group was significantly more accurate on the ‘mmms’ task ($Z = 2.74, n = 25, p < 0.01$). On the faces task, the comparison group was also more accurate at identifying emotion from the face ($t = 4.197; df = 23, p < 0.001$).

The difference between the ASC group and the comparison group in the incongruent and neutral conditions and the lack thereof in the congruent condition indicate that the participants with ASC were generally relying more on the semantics of the sentence in identifying the emotions in the stimulus sentences. If this was the case, we would also expect their errors in the incongruent trials to tend towards the emotion represented by the sentence semantics (which did not match with the emotion encoded in the voice). This was indeed the case. In the incongruent trials, people with ASC selected a semantically matching emotion significantly more frequently than people from the comparison group (ASC: mean = 0.51, SD = 0.23; comparison group: mean = 0.28, SD = 0.12; $t = 3.297, df = 23, p < 0.01$).

To assess whether performance on the various tests was related, correlations were taken between the voice emotion recognition task and the facial emotion task, as well as the Mill Hill test (Table 2). The faces task was correlated with the incongruent and neutral conditions of the sentence-based task, but not with the ‘mmms’ task. None of the tasks were significantly correlated with performance on the Mill Hill vocabulary test.

<table>
<thead>
<tr>
<th>Table 1. Performance on voice, faces and verbal ability tasks by group.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ASC Group</strong></td>
</tr>
<tr>
<td>Mean (SD)</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Accuracy congruent emotion</td>
</tr>
<tr>
<td>Accuracy incongruent emotion</td>
</tr>
<tr>
<td>Accuracy neutral emotion</td>
</tr>
<tr>
<td>‘mmms’ accuracy</td>
</tr>
<tr>
<td>Faces accuracy</td>
</tr>
<tr>
<td>Mill Hill (verbal IQ)</td>
</tr>
<tr>
<td>Age</td>
</tr>
</tbody>
</table>

Between-group comparisons: **p < 0.001*, p < 0.005.
Discussion

This study reports on a new test of vocal emotion, and tested whether individuals with ASC showed different degrees of divergence from typically developed individuals in interpreting emotions from the voice. We predicted that individuals with ASC would identify emotions at a similar level of performance to a typically developed comparison group when the intended emotion in the voice was also discernable from the semantic content of the sentence. Indeed identification of emotion in read sentences did not differ between the ASC group and the comparison group when the emotion in the voice and the meaning of the sentence were congruent. However, as predicted, identification was significantly poorer in people with ASC for the trials when the sentence meaning and emotion in the voice were incongruent, when the meaning of the sentence was neutral, and for the trials when there was no lexical or semantic information (the ‘mmms’).

Second, we predicted that difficulties in interpreting emotions from voice and faces would be closely related in individuals with ASC, because both are subjected to a single cause in general cognition rather than modality-dependent problems in cue processing. There were significant correlations between emotion recognition from voice and from faces although the results must be treated with caution as the sample size is small. In addition, the correlation between ‘mmms’ and faces was not statistically significant. This may be due to the nature of the task, and that in general typically developing individuals found this task more difficult, with much lower accuracy.

There are a few important implications to be drawn from the results of the voice emotion recognition task, in which the ASC group, who otherwise found the task difficult, did not differ from
the comparison group when there was matching information between the prosody and the semantics of the sentences. Indeed, the participants with ASC were the least accurate on the sentence task in which the emotion in the voice and meaning of the sentence were incongruent, showing a much larger drop in accuracy than the comparison group participants. In real-life exchanges we would expect that vocal emotion would be congruent with the semantic content of an utterance and, when this is not the case, that the message would be judged by the vocal emotion rather than the message content (Mehrabian and Wiener, 1967). An exception to congruent vocal emotion and semantic content are irony and sarcasm, where this congruency is flouted to give rise to a non-literal interpretation of the utterance. The current study thus suggests that the impairment in understanding irony in ASC (Wang et al., 2006) may be due to a difficulty using the vocal information to override the semantics of an utterance. In this regard, it is also interesting to note that people with ASC are often anecdotally reported to have a preference for communicating via email rather than in person.

The results of our study are consistent with the findings from Grossman et al. (2000) and Lindner and Rosén (2006), who also found ASC individuals’ tendency to rely on semantic information in decoding emotion. When no semantic information was given, the individuals with ASC had more difficulty in identifying the emotion in the voice. In addition, when the semantic information was conflicting with the prosodic information (in the incongruent trials) individuals with ASC selected a semantically matching emotion significantly more than people from the comparison group. Individuals with ASC may use a range of strategies to understand affect in others, and may need to decode the information in a different way to typically developed individuals who tend to process emotional information in a more ‘automatic’ manner, using facial and prosodic cues. Whether these strategies mean that the individual understands what it is for another to have an emotion is beyond the scope of this study.

The participants in the current study and those in Grossman et al.’s (2000) and Lindner and Rosén’s (2006) were all high functioning. Grossman et al. (2000) included only individuals with AS and those whose full-scale IQ was greater than 80 and verbal IQ greater than 90 on the Wechsler Intelligence Scale for Children, Third Edition (Wechsler, 1992). The current study included only those with HFA or those with AS. It is likely that the strategy of using verbal information rather than prosody or faces may not apply in groups whose IQ is lower or whose verbal skills are not as developed.

This study has a number of further limitations. It is a very small study, and how much these findings can be generalized to the ASC population as a whole is unknown. Indeed, as discussed, the findings may be applicable only to high functioning groups. Within that caveat, there are reasons to believe that our participants constitute a representative sample of the ASC population. Apart from the fact that all of the clinical participants were identified with HFA or AS (according

### Table 2. Correlations between the emotional recognition measures and the Mill Hill (verbal IQ) test.

<table>
<thead>
<tr>
<th></th>
<th>Faces</th>
<th>Mill Hill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congruent</td>
<td>0.25</td>
<td>-0.15</td>
</tr>
<tr>
<td>Incongruent</td>
<td>0.55**</td>
<td>0.39</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.44*</td>
<td>0.13</td>
</tr>
<tr>
<td>'mms'</td>
<td>0.18</td>
<td>0.11</td>
</tr>
<tr>
<td>Faces</td>
<td>–</td>
<td>0.27</td>
</tr>
</tbody>
</table>

**p < 0.01; *p < 0.05
to DSM criteria), the outcomes are consistent with other studies that assess emotional processing (Balconi and Carrera, 2007; Dyck et al., 2001; Humphreys et al., 2007; Peppé et al., 2007).

In conclusion, our study shows that despite difficulties shown in interpreting emotions from voice and faces, individuals with ASC can use semantic strategies to identify emotions. There is a shared mechanism behind problems in recognizing emotions in other individuals regardless of the modality of the information (acoustic-prosodic versus visual-facial). Although it remains to be tested whether such difficulties may be related to their comprehension of others’ intentions, they are consistent with a model of ASC that places the locus of the deficit in the ability to infer emotional states in others.

Acknowledgements
We thank all the participants who took part, and the staff of Number 6.

Funding
This work was supported by the British Academy (SG_49424).

References


