PRODUCTION AND PERCEPTION OF AFFECTIVE PROSODY
BY ADULTS WITH AUTISM SPECTRUM DISORDER

by

Daniel J. Hubbard

APPROVED BY SUPERVISORY COMMITTEE:

___________________________________________
Dr. Peter F. Assmann, Chair

___________________________________________
Dr. Carlos Busso

___________________________________________
Dr. William F. Katz

___________________________________________
Dr. Noah J. Sasson
Dedicated to my daughters, Madison and Alison
PRODUCTION AND PERCEPTION OF AFFECTIVE PROSODY
BY ADULTS WITH AUTISM SPECTRUM DISORDER

by

DANIEL J. HUBBARD, BA, MS

DISSEPTION
Presented to the Faculty of
The University of Texas at Dallas
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY IN
COGNITION AND NEUROSCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

December 2016
ACKNOWLEDGMENTS

First, I would like to thank my family and friends for their support of my educational journey. In particular, I thank my parents, Ron Hubbard and Linda Hesse, for always believing in me, even when the future was unclear. This accomplishment would not have been possible without their unconditional love and support. I also thank my wife, best friend and soul mate, Kary, for her support and for helping to push me along when progress seemed slow. I thank my daughters for being my biggest fans. Madison and Alison, my hope for you is that you always believe in yourselves and have confidence that you can achieve amazing things as long as you keep your dreams alive and work hard to make progress toward your goals each day. I love you dearly.

I thank my brothers, Jason and Paul Hubbard, and cousins Corey Jones, Clay Jones, Curtis Jones, and Christopher Jones, and their families, and many friends for your interest in my research, even when you were pretending, and for providing much needed fun along the way.

I would also like to sincerely thank and acknowledge Dr. Peter Assmann, my research advisor, greatest mentor, and friend. Dr. Assmann was pivotal in my decision to pursue cognitive science and the graduate programs at UT Dallas. From the time I first walked into his classroom, I have had deep respect and admiration for his passion for teaching and research in cognitive science and speech production and perception. Thanks also to my supervisory committee members, Professors Carlos Busso, William Katz, and Noah Sasson. The training you’ve provided and the conversations we’ve had about this project have been invaluable.
I also thank members of the Assmann Lab and the graduate program in Cognition and Neuroscience at UT Dallas. Specifically, Shaikat Hossain, Vahid Montazeri, Michelle Kapolowicz, Olga Peskova, Adrian Cho, Daniel Guest, Daniel Faso, Kerrianne Morrisson, and Professors Bert Moore, James Bartlett, Hervé Abdi and Alice O’Toole were instrumental during my training.

Finally, special thanks and gratitude is extended to the individuals who took part in this project and their families, and to the Nonpareil Institute of Plano, TX for helping with participant recruitment.

October 2016
Affective prosody, defined as the use of paralinguistic elements in speech to convey emotion, is important for effective social functioning. While generally a trivial task for typically-developing (TD) adults, individuals with autism spectrum disorder (ASD) present with significant challenges in social communication and interaction, including prosody. Previous research has shown that talkers with ASD produce pragmatic prosody with increased variability in fundamental frequency (f0, which is closely correlated with voice pitch), but it was unclear whether those differences carry over to speaking tasks involving emotion elicitation. A controlled set of expressive speech recordings was obtained from talkers with ASD and controls in five emotion contexts: angry, happy, interested, sad, and neutral. Emotion-specific group differences in f0, intensity, and duration were found in multiple speech types, and the pattern of results was characterized by inconsistent and exaggerated patterns of affective prosody production in talkers with ASD compared to controls. The perceptual relevance of the affective acoustic differences
was tested in three listening experiments involving talkers and listeners with ASD and controls. The first two experiments involving TD listeners were designed to examine the perceptual impact of increased f0 variability and intensity found in recordings produced by talkers with ASD. Compared to the intensity manipulation, modifying the f0 contour had a larger impact on emotion recognition accuracy. The third experiment was designed to compare perception of affective prosody in listeners with ASD and controls using unmodified stimuli, and revealed that differences in affective prosody perception were more closely related to talker group production differences than listener group differences. The results are consistent with previous work in face perception showing increased emotion identification rates but lower naturalness ratings when listeners responded to stimuli produced by individuals with ASD. The findings are interpreted within the context of the speech attunement framework, which suggests that individuals with ASD lack the motivation to attune their prosodic speech to sound like TD talkers.
TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... v

ABSTRACT ............................................................................................................................... vi

LIST OF FIGURES .................................................................................................................. ix

LIST OF TABLES ...................................................................................................................... x

CHAPTER 1: INTRODUCTION ................................................................................................. 1

1.1 Prosody ............................................................................................................................... 2

1.2 Affective prosody ................................................................................................................. 4

1.3 Autism spectrum disorder .................................................................................................. 6

1.4 Communication of affective prosody in individuals with ASD ....................................... 7

1.5 A brief overview of research conducted ............................................................................. 11

CHAPTER 2: AFFECTIVE PROSODY PRODUCTION .............................................................. 16

2.1 Introduction ......................................................................................................................... 16

2.2 Method and procedure ....................................................................................................... 20

2.3 Innovation .......................................................................................................................... 26

CHAPTER 3: ACOUSTIC ANALYSIS OF AFFECTIVE PROSODY ........................................ 28

3.1 Introduction ......................................................................................................................... 28

3.2 Method and procedure ....................................................................................................... 30

3.3 Hypotheses ......................................................................................................................... 32

3.4 Innovation .......................................................................................................................... 33
3.5 Results ........................................................................................................................................... 34
3.6 Functional data analysis .................................................................................................................. 48
3.7 Discussion ....................................................................................................................................... 60

CHAPTER 4: AFFECTIVE PROSODY PERCEPTION ............................................................................. 65

4.1 Introduction ...................................................................................................................................... 65
4.2 Experiment 1: The role of f0 in affective prosody perception ....................................................... 69
4.3 Experiment 2: The role of intensity in affective prosody perception ............................................. 81
4.4 Experiment 3: Affective prosody perception in listeners with ASD ............................................ 90
4.5 Modeling perception of affective prosody ....................................................................................... 100

CHAPTER 5: GENERAL DISCUSSION AND CONCLUSION ............................................................. 106

REFERENCES ......................................................................................................................................... 111

VITA
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Dimensions of human affective response: arousal level and valence</td>
<td>12</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>Workflow diagram</td>
<td>15</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Expressive speech database recordings protocol</td>
<td>26</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>F0 range (in Hz) for evoked emotional phrases produced by talkers with ASD and TD controls</td>
<td>35</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>F0 contours for individual evoked phrases and 95% confidence bands (superimposed in black) by talker group for each emotion context and repetition for the phrase “why did you do that?”</td>
<td>36</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Mean intensity levels (dB SPL) for evoked phrases produced by talkers with ASD and TD controls</td>
<td>37</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Mean phrase duration (in ms) for evoked phrases produced by talkers with ASD and TD controls</td>
<td>39</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>F0 range (in Hz) for VCV syllables produced by talkers with ASD and TD controls</td>
<td>41</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Mean intensity levels (dB SPL) for VCV syllables produced by talkers with ASD and TD controls</td>
<td>42</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>Mean intensity contours for VCV syllables produced by talkers with ASD and TD controls</td>
<td>43</td>
</tr>
<tr>
<td>Figure 3.8</td>
<td>Mean duration (in ms) for VCV syllables produced by talkers with ASD and TD controls</td>
<td>44</td>
</tr>
<tr>
<td>Figure 3.9</td>
<td>Mean f0 range (in Hz) for scripted dialogue recordings produced by talkers with ASD and TD controls</td>
<td>46</td>
</tr>
<tr>
<td>Figure 3.10</td>
<td>Mean intensity (in dB SPL) for scripted dialogue recordings produced by talkers with ASD and TD controls</td>
<td>47</td>
</tr>
</tbody>
</table>
Figure 3.11  Mean phrase duration (in ms) for scripted dialogue recordings produced by talkers with ASD and TD controls ................................................................. 48

Figure 3.12  Functional principal component analysis results (principal component 1) for productions of the evoked phrase “why did you do that” produced in the happy emotion context ................................................................. 53

Figure 3.13  Example f0 contour FDA plot for the evoked phrase “why did you do that” produced in a happy emotion context ................................................................. 54

Figure 3.14  Functional principal component analysis results (principal component 1) for productions of the evoked phrase “I can’t believe this” produced in the angry emotion context ................................................................. 58

Figure 3.15  Example f0 contour FDA plot for the evoked phrase “I can’t believe this” produced in the happy emotion context ................................................................. 59

Figure 4.1  Example f0 contours for the evoked phrase “what, what do you mean?” produced in the happy context ................................................................. 73

Figure 4.2  Listener emotion recognition accuracy (ERA) for evoked phrases produced by talkers with ASD and TD controls ................................................................. 77

Figure 4.3  Mean listener naturalness ratings for evoked phrases produced by talkers with ASD and TD controls ................................................................. 78

Figure 4.4  Mean listener emotion recognition accuracy (ERA) scores for evoked phrases produced by talkers with ASD and TD controls ................................................................. 86

Figure 4.5  Mean listener naturalness ratings for evoked phrases produced by talkers with ASD and TD controls ................................................................. 87

Figure 4.6  Mean listener emotion recognition accuracy (ERA) scores for evoked phrases produced by talkers with ASD and TD controls ................................................................. 96

Figure 4.7  Mean listener naturalness ratings for evoked phrases produced by talkers with ASD and TD controls ................................................................. 97

Figure 4.8  Predicted neutral responses based on variation in f0 (left panel) and intensity (right panel) ................................................................. 101

Figure 4.9  Predicted angry responses based on variation in f0 (left panel) and intensity (right panel) ................................................................. 102

Figure 4.10  Predicted happy responses based on variation in f0 (left panel) and intensity (right panel) ................................................................. 103
Figure 4.11  Predicted interested responses based on variation in f0 (left panel) and intensity (right panel)……………………………………………………………………………… 103

Figure 4.12  Predicted sad responses based on variation in f0 (left panel) and intensity (right panel)……………………………………………………………………………… 104
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Example dialogue script for the happy emotion context</td>
<td>24</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Evoked phrase acoustic data: group mean, standard deviation, and range for f0, intensity, and duration by emotion context and talker group</td>
<td>34</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>VCV acoustic data: group mean, standard deviation, and range for f0, intensity, and duration by emotion context and talker group</td>
<td>40</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Scripted dialogue acoustic data: group mean, standard deviation, and range for f0, intensity, and duration by emotion context and talker group</td>
<td>45</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Mean functional principal component scores and standard deviations for the analysis of f0 contours extracted from evoked phrases</td>
<td>52</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Mean functional principal component scores and standard deviations for the analysis of intensity contours extracted from evoked phrases</td>
<td>57</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

Human speech is a complex and versatile communication mode consisting of at least two streams of information – the linguistic stream, which is reserved for transactional purposes (e.g. “what is said”), and the paralinguistic stream (e.g., “how speech is said”), which is reserved for socio-cultural information (Brown & Yule, 1983). The research described in this dissertation examined communication of paralinguistic content in speech involving emotional expression, known as affective prosody. Impairments in affective prosody communication are a core symptom of autism spectrum disorder (ASD) for autistic individuals who speak, and there is a strong correlation between social communication difficulties and prosodic abnormalities in ASD (Paul et al., 2005). In this project, a series of experiments was conducted to investigate production and perception of affective prosody in adults with ASD and typically-developing (TD) controls. The primary goals for this project were to quantify acoustic differences in affective prosody production between the talkers with ASD and TD talkers serving as controls, and to measure the perceptual impact of those differences for listeners.

The current document is divided into five chapters, consisting of an introduction (Chapter 1) followed by three chapters covering separate major phases of work: Chapter 2 describes the database of emotional speech recordings collected from talkers with ASD and TD controls; Chapter 3 details acoustic analysis and statistical modeling conducted to assess affective prosody production differences between the two groups; Chapter 4 describes perceptual tests and results
using modified stimuli designed to examine perception of affective prosody in ASD and TD listeners; Chapter 5 is a summary and discussion of the experiments and results obtained throughout. The remainder of the introductory chapter is organized as follows: prosody and affective prosody are defined and discussed in Section 1.1; a literature review of research on communication of affective prosody in TD populations is presented in Section 1.2; a brief overview of ASD is presented in Section 1.3; research concerning affective prosody communication in individuals with ASD is discussed in Section 1.4; and finally, the series of experiments designed to examine affective prosody production and perception in ASD is introduced in Section 1.5.

1.1 Prosody

The term “prosody” is used to describe the use of suprasegmental elements in speech, e.g., properties that span multiple sound segments, to convey meaning (Lehiste, 1970). Suprasegmental properties contrast with segmental elements, consisting of individual vowels and consonants, which comprise the basic building blocks of phonemes and words. Dating back several thousand years, early Greek and Roman rhetoric manuals contained examples of voice manipulations used to convey different emotional states (Kennedy, 1972). The Greek word “προσῳδία” (prosodia), the origin of the term “prosody,” was used to refer to the rhythm, stress, and intonation of speech. Thus, the ability to convey meaning using prosody has been studied for thousands of years.

Suprasegmental and segmental elements are transmitted concurrently in a given utterance, and while the linguistic component is left intact, talkers retain the ability to change the meaning of the utterance using systematic modulations to suprasegmental voice properties.
Modulations in fundamental frequency (f0, which is correlated with perceived voice pitch), intensity, and the type and duration of silences compared to vocalic portions of an utterance are used to signal a talker’s intended meaning (Kent & Read, 2002). These modulations are also used by listeners as cues for perceiving a talker’s internal affective state (Banse & Scherer, 1996). Due to its role in shaping speech communication beyond what is conveyed by words alone, prosody is regarded as a vital aspect of speech perception and social functioning (Banse & Scherer, 1996; Scherer, 2003).

The larger domain of prosody consists of three more specific types. First, grammatical prosody conveys syntactic information such as stress patterns that signal the meaning of words (Shriberg et al., 2001). For example, the meaning of the word “suspect” differs depending on whether stress is placed on the first or second syllable by the talker (Paul el al., 2005). With the capitalized portion of the word indicating more stress, “SUS-pect” is a noun that refers to someone accused of an offense, whereas “sus-PECT” is a verb that means to have an idea about the truth of something. Second, pragmatic prosody conveys social information such as the use of stress to direct the listener to the focus of attention within a sentence. For example, the meaning of the sentence “are you there yet” differs depending on whether stress is placed on the word “there” (to inquire about a conversation partner’s location) or the word “yet” (perhaps to convey that someone is late). Finally, affective prosody is reserved for emotional information, such as conveying a talker’s internal affective state. Affective prosody is discussed in detail in section 1.2 below.
1.2 Affective prosody

Emotion and speech are at the very core of human social life. Even before infants are able to express emotion using language, they convey their affective state vocally using cries, coos and other precursors to speech. As infants develop and learn language, the intrinsic relationship between emotion and speech flourishes. From that point throughout adulthood, speech is one of the primary methods by which one’s affective state is communicated with others, and indeed, it is the only method available when visual cues are absent (e.g., phone conversations). The term affective prosody will be used throughout this dissertation to refer to the use of suprasegmental features in speech to convey emotion. The social relevance of affective prosody is highlighted by the fact that the perceived meaning of an utterance can be altered by the use of different patterns of prosodic cues. For example, even in short utterances such as the word “hello,” the talker’s intended meaning can be altered using different patterns of acoustic modulations to convey happiness (rising pitch contour, longer speaking rate) or anger (falling pitch contour, shorter speaking rate).

Research on affective prosody began in earnest in the 1960s and 1970s, accompanied by an increase in acoustic analysis and perceptual tests involving emotional speech recordings (Lieberman & Michaels, 1962; Scherer, 1979, 1986). This body of research revealed systematic voice property modulations used in production of different emotion categories (Scherer, 1979, 1986; Murray and Arnott, 1993; for a review see Juslin and Laukka, 2003). Much of this work examined f0, intensity and speaking rate, which are generally correlated with one’s physiological state of arousal (Banse & Scherer, 1996). For example, high arousal emotions (e.g., happiness and anger) are accompanied by increases in f0 range, f0 mean, speaking rate, and intensity,
whereas low arousal emotions such as sadness are characterized by decreased f0 range, f0 mean, speaking rate, and intensity.

Recent studies have broadened the research focus to include voice quality parameters such as spectral slope and harmonics-to-noise ratio (Goudbeek & Scherer, 2010). F0 is a carrier of emotional information in speech that aids in signifying the physiological arousal level of the talker, while voice quality measures have been linked to the valence dimension, which corresponds perceptually to how positive or negative the speech sounds (Goudbeek & Scherer, 2010; Waaramaa et al., 2010). Valence and arousal level are widely regarded as basic emotion dimensions (Wundt, 1909; Lang et al., 1993; Mehrabian & Russell, 1974). The valence dimension corresponds to how attractive (positive valence) or aversive (negative valence) an object of emotion is perceived to be (Wundt, 1909; Russell, 1980, 2003; Goudbeek & Scherer, 2010), whereas arousal corresponds to a general physiological level of excitement or engagement (Frijda, 1986; Goudbeek & Scherer, 2010).

Russell (1980) proposed a “circumplex” model of emotional response in which valence and arousal are primary bipolar dimensions. The model was based on perceptual ratings from participants using multiple scaling methods. In one task, the relevance of the arousal and valence dimensions was tested by asking participants to order a series of eight emotion terms around a circle. Ten of the 36 participants ordered the emotion terms in the precise order predicted by the model, with “aroused” and “sleepy” occupying opposite ends of the arousal dimension and “pleasure” and “misery” occupying opposite ends of the valence dimension. The modal response for each emotion term was congruent with the predictions made by the theoretical model (Russell, 1980). In speech, converging evidence has suggested that the acoustic and perceptual
differences between emotion categories can often be linked to the valence and arousal level of the talker (Cowie & Cornelius, 2003; Juslin & Laukka, 2003; Banziger, Mortillaro, & Scherer, 2012; Goudbeek & Scherer, 2010). A current area of research interest involves the search for reliable acoustic cues related to the valence dimension (Busso & Rahman, 2012). While the role of f0 and its impact on arousal level in speech has been established for some time, understanding voice cue modulations characteristic of the valence dimension has been more elusive.

1.3 Autism spectrum disorder

Autism is a neurodevelopmental disorder characterized by deficits in social interaction, verbal and nonverbal communication, and repetitive behaviors. This pattern of symptoms, which manifest in varying degrees in an autistic person, is known commonly as autism or autism spectrum disorder (ASD). Symptoms of ASD include a wide range, or “spectrum” of deficits, abnormal behaviors and skills that vary in occurrence and severity from one individual to the next. Diagnosis occurs through behavioral evaluations conducted by trained physicians. Several instruments such as the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 1989) and Autism Diagnostic Interview (ADI/ADI-R; Le Couteur et al., 1989; Lord, Rutter, & Le Couteur, 1994) are commonly used standardized assessments used by physicians for diagnosis, which can occur as early as 18 months of age or younger. Often, parents are first to notice abnormal behaviors such as the child failing to respond to their name when called, failing to make eye contact, or unusual, repetitive behaviors.

The beginning of modern scientific research on autism occurred in 1943, when Leo Kanner published a seminal paper describing the symptoms of 11 children with a similar group of behaviors that had not been previously identified (Kanner, 1943). Kanner’s influential report
marked the first thorough description of autism as a disorder rather than as a set of symptoms. The 11 children in Kanner’s report included 8 boys and three girls between 2 to 8 years of age that all shared high intelligence, yet had language deficiencies, monotonous repetitive behaviors, and an intense desire to maintain sameness. Since that time, diagnosis of ASD has increased dramatically, partly due to changes in diagnostic criteria and increased public awareness about the disorder. The Centers for Disease Control (CDC) reported an increase in the prevalence of ASD to approximately 1 in 68 children, according to 2012 estimates from CDC’s Autism and Developmental Disabilities Network (Christensen, et al., 2016).

A key deficiency included in the diagnostic criteria for ASD listed in the Diagnostic and Statistical Manual of Mental Disorders (5th. Ed.; DSM-5) is disordered social communication and interaction, often exhibited by a “lack of social or emotional reciprocity” (American Psychiatric Association, 2013). Emotional reciprocity involves the ability to accurately perceive emotional content communicated by others, and to respond appropriately to that content. It is precisely that part of the spectrum of deficits that ASD presents that this thesis was designed to investigate, as the research described here compared production and perception of affective prosody among adults with ASD and a group of non-autistic controls.

1.4 Communication of affective prosody in individuals with ASD

While non-autistic adults normally utilize affective prosody without difficulty, a core feature of ASD is disordered social communication, including abnormal prosodic functioning. Abnormal production of prosody is socially apparent, and prosodic abilities are routinely assessed in widely used diagnostic tools for ASD (e.g., ADOS; Lord et al., 2000), yet research that has objectively measured differences in prosodic functioning between adults with ASD and
TD controls is still somewhat limited. Current diagnostic methods for ASD related to affective prosody involve qualitative physician- or caregiver-administered assessments such as the Prosody-Voice Screening Profile (Shriberg, Kwiatkowski, & Rasmussen, 1990) and Profiling Elements of Prosodic Systems-Children (Peppé & McCann, 2003). There is an absence of a quantitative method for assessing affective speech production deficits. The research described in this thesis may enhance objective methods for assessing prosodic functioning by quantifying differences in affective prosody production and perception among adults with ASD and TD controls.

Research in this area has shown that individuals with ASD exhibit greater variation in f0 compared to control participants (Fosnot & Jun, 1999; Nadig & Shaw, 2012). Fosnot and Jun (1999) compared sentences spoken by four autistic children and four controls in two conditions – one in which the sentences were read by the participants and a second in which the sentences were repeated after the experimenter read them (imitation). The autistic children exhibited increased f0 variability (e.g., mean f0 range) and longer sentence durations compared to the control group in both the reading and imitation conditions. In addition, autistic children produced the sentences with greater variability in f0 and sentence duration compared to the controls.

Nadig and Shaw (2012) compared conversational speech and structured communication samples collected from 15 children with ASD and 13 age-matched control participants who ranged in age from 8 to 14 years. Consistent with the earlier results of Fosnot and Jun (1999), acoustic analysis of the recordings revealed increased f0 range in talkers with ASD. Nadig and Shaw (2012) also presented the recordings to listeners and asked them to rate how typical they were. While listeners reported atypical prosody in the ASD group, and acoustic analysis revealed
greater f0 variation, listeners did not rate the ASD talkers as producing greater f0 variation. The authors concluded that the f0 variation displayed by their talkers with ASD was unconventional and therefore was not perceived as f0 variation by their listeners. Increased f0 variation in talkers with ASD has also been reported in languages other than English. Consistent with the findings using English-speaking talkers, Sharda and colleagues (Sharda, et al., 2010) and Green and Tobin (2009) reported increased f0 range in speech produced by ASD talkers in Hindi and Hebrew, respectively.

Research has also shown that disordered affective prosody communication can lead to perceptions of “social oddness” regarding speech produced by participants with ASD (Van Bourgondien & Woods, 1992; Paul et al., 2005). In Paul et al. (2005), correlations were found between voice deficits and ratings of social and communicative abilities in talkers with ASD. Participants with ASD who displayed high scores on measures of the appropriate use of prosody were more likely to receive better scores on overall communication and socialization ratings.

It is important to note, however, that not all adults with autism show deficits in communication of affective prosody, and individuals affected by ASD often improve their prosody production and perception abilities throughout the course of development. Given the heterogeneous nature of ASD, it is still unclear to what extent individuals with ASD share deficits in affective prosody functioning. In a study that assessed affective prosody production in 30 children and adults ranging in age from 10-49 years, only half of the participants were reported as having deficits (Shriberg et al., 2001). However, in many individuals with ASD, deficits persist into adulthood and are often not targeted in clinical treatment methods (McCann et al., 2007; Paul et al., 2005).
The results of studies to date in this area represent an apparent contradiction with the earliest clinical accounts of autism (Kanner, 1943). Kanner reported a mechanical speaking style and “monotonously repetitious” verbal utterances common in his case studies. An important consideration, however, is that Kanner’s descriptions included moderate to severely autistic individuals, whereas the recent research has been conducted using ASD participants with milder forms of impairment. The participants with ASD in Nadig and Shaw (2012) were considered high-functioning, with a mean performance IQ of 105 (range: 81-126) for the conversational speech task and 111 (range: 81-133) for the structured communication task (measured using WASI); Fosnot & Jun (1999) did not report performance IQ scores for their four child/adolescent ASD participants but mentioned they were sight-word readers and in special education classes in school. In contrast, the case studies reported by Kanner (1943) included more severe examples of autism.

The research described in this thesis may be helpful for enhancing objective diagnostic and treatment methods for ASD involving affective prosody. The studies described above represent some of the first work to identify affective prosody differences between individuals with ASD and TD controls; however, additional research is needed using controlled speech recordings and emotion elicitation to better understand divergent affective prosody production and perception in individuals with ASD, and the perceptual impact of abnormal production on emotion perception. The research described in this dissertation used a controlled set of speech recordings collected for acoustic analysis and perceptual testing (see Chapter 2). Previously published work in this area has been somewhat limited by studies involving small samples of talkers and few studies using adults with ASD.
1.5 A brief overview of research conducted

The research described in this thesis examined affective prosody communication in a group of adult males with ASD, and compared production and perception of affective prosody to a TD control group. ASD is characterized by deficits in social interaction and communication, including emotional reciprocity, which refers to normal back-and-forth conversation related to sharing of interests, emotions or affect (American Psychiatric Association, 2013). A comprehensive set of speech recordings was collected in which participants with ASD and TD controls produced an identical set of utterances in five emotion contexts: angry, happy, interested, sad, and neutral. The contexts were chosen because they represent distinct locations on the acoustic space defined by the valence and arousal dimensions: happy (high arousal/positive valence); interested (low arousal/positive valence), angry (high arousal/negative valence), and sad (low arousal/negative valence). Neutral recordings were collected for comparative purposes. Figure 1.1 displays the valence and arousal dimensions and the location of selected emotion categories along those dimensions.

The recordings consisted of short ambiguous phrases, nonsense vowel-consonant-vowel (VCV) syllables, and scripted dialogue sessions. Some previous work has used spontaneous speech samples for comparison between ASD and TD talkers; however, the research reported here involved a controlled set of utterances in which talkers portrayed the five emotion contexts while producing the same lexical content. The short ambiguous phrases contained neutral lexical content, meaning that each phrase could be realistically produced in any of the five emotion contexts. The VCV syllables were chosen to eliminate the potential impact of prior experience with the target speech on production, and the scripted dialogue sessions enabled analysis of
affective prosody in utterances with matching emotion and lexical content. The set of recordings provided data for acoustic analysis, and the short ambiguous phrases were used in listening experiments designed to test hypotheses relevant to specific voice cues involved in affective prosody perception.

![Image of a circumplex model of affect](image)

Figure 1.1. Dimensions of human affective response: arousal level and valence. Reproduced based on Russell, J. (1980). “A circumplex model of affect” Journal of Personality and Social Psychology, 39, 1161-1178 (Figure 3).

The recordings were obtained using two elicitation techniques utilized in previous research: evoked and prompted portrayals. In the evoked method, talkers were asked to produce the target utterances while recalling an emotional episode in their past that matched the target emotion context. In the prompted method, talkers were motivated using an expressive prompt designed to help them produce the target emotion context. For example, an expressive prompt used for the angry context involved a brief story about lost luggage at the airport. The two
elicitation methods allowed comparison of affective prosody produced under separate ecological
criteria: imagined emotional experiences (portrayed) and those that are remembered from
past experience (evoked). Previous research in the face perception literature has suggested that
measurable differences exist between evoked and prompted facial expressions and that evoked
(or “felt”) expressions are most closely linked to expected patterns of changes in faces (Gosselin,
Kirouac, & Doré, 1995). Both elicitation methods were used to determine if the previously
observed differences in evoked and prompted facial expressions are also present in speech, and if
so, whether those differences exist in ASD talkers. See Chapter 2 for more details about the
obtained speech recordings. See Figure 1.2 for a work-flow diagram of the recordings protocol
and other steps involved in this project.

The recordings were processed using measurement algorithms designed to estimate and
extract acoustic properties that are potentially relevant for production and perception of affective
prosody. Each recording was divided into non-overlapping 1 ms segments prior to obtaining the
measurements. Acoustic analyses were performed on the measurements to assess group
differences between talkers with ASD and TD controls, and to identify acoustic properties
potentially used by listeners during perception of affective prosody.

Listening tests were performed in which a subset of the recordings was presented to naïve
participants who were asked to identify the emotion context presented and rate the level of
“naturalness” in each recording. Naturalness was defined as the extent to which the emotional
content in each test stimulus would be encountered in everyday expressive communication. The
perceptual data was an important addition used to model the link between production and
perception of affective prosody, and differences between the ASD and TD groups. Statistical
analysis methods included analysis of variance (ANOVA), functional data analysis (Ramsay & Silverman, 2005) and regression methods. ANOVA was used to identify group differences in affective prosody production between ASD and TD talkers, and functional data analysis was used to identify detailed differences in patterns of f0 and intensity production. Using the regression approach, listening test results were modeled using the acoustic measurements to identify patterns of emotion-specific changes in the acoustic properties that were perceptually relevant for listeners. Additional details and results of the acoustic analysis can be found in Chapter 3.

The perceptual relevance of speech properties identified in the acoustic analysis was examined by conducting listening tests using original and synthesized versions of the database recordings. The goal of the listening tests was to test the relevance of candidate cues for affective prosody perception using systematically altered versions of the original recordings, and to compare perception of affective prosody in listeners with ASD and controls. Test stimuli were created using recordings in which the potentially relevant property was modified while holding all other properties constant. For example, previous work has shown that the shape of the f0 contour can be an important cue for affective prosody perception, given that happy speech has shown greater f0 modulation, as measured by increases in f0 mean and f0 range, compared to other contexts. Therefore, the f0 contour was modified in Experiment 1 to determine the perceptual impact of changes in f0 on perception. An additional manipulation was performed on the intensity contour of the same group of test stimuli. See Chapter 4 for additional information and results of the experimental manipulations tested.
A theoretical basis for the observed deficits is that individuals with ASD exhibit disordered social attention functioning compared to TD individuals. Research has suggested that individuals with ASD are less likely to pay attention to salient social stimuli compared to control subjects, a phenomenon referred to as diminished social attention (Klin et al., 2003). Inattention to social norms related to emotion is a likely contributor to differences in prosody production, due to the inability to utilize those norms during speech production. A prediction based on the diminished social attention hypothesis was that talkers with ASD would exhibit greater variability in production of acoustic properties relevant to affective prosody communication compared to TD talkers. The research conducted represents a comprehensive study of differences in affective prosody communication between ASD and TD participants, and has the potential to enhance diagnostic tools and training paradigms designed to help individuals with ASD develop socially acceptable forms of affective prosody communication.

Figure 1.2. Work flow diagram to examine production and perception of affective prosody in adults with autism spectrum disorder (ASD) and typically-developing (TD) controls
CHAPTER 2
AFFECTIVE PROSODY PRODUCTION

2.1 Introduction

Human speech communication is fascinating in its complexity. Considering only the linguistic stream of information, speech production requires that a limited set of language-specific speech sounds are used in a meaningful way to form larger units of speech such as words and sentences. There is an added layer of complexity arising from the nonlinguistic stream, based on the fact that a single word or phrase can be produced in a variety of ways to convey different meanings. Due to its ability to convey meaning beyond that which is carried by the lexical content alone and to modulate and enhance meaning, prosody has been called “the music of speech” (Wennerstrom, 2001). This chapter reports on the collection of a database of expressive speech recordings designed to examine affective prosody production and perception differences in adult males with ASD and a group of TD control participants.

Temporal modulations in emotionally-salient voice properties are used to signal a talker’s affective state. For example, the production of angry speech is often accompanied by increased intensity and shorter utterance durations compared to other emotions. Within the nonlinguistic stream of information that accompanies any utterance, affective prosody contrasts the two other forms of prosody – grammatical prosody, which is used to signal syntactic information such as whether an utterance is a question or a statement, and pragmatic prosody such as contrastive
stress that helps a listener identify information that is new to the conversation (Shriberg et al., 2001).

Within the TD population, studies of affective prosody have revealed systematic differences in acoustic properties associated with individual emotion categories (Scherer, 1979, 1986; Murray & Arnott, 1993; for a review, see Juslin and Laukka, 2003). Much of this work has investigated modulations in f0, intensity, 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). Those differences are known to be important for perception of affective prosody (Tóth, Sztahó, & Vicsi, 2008; Waaramaa et al., 2010). For example, 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).

For individuals with ASD, however, pervasive social disability manifests not only in social perception and cognition (Sasson et al., 2011) but also in differences in social expressivity (Begeer et al., 2008), including atypical facial affect (Faso, Sasson & Pinkham, 2015) and speech production (Fosnot & Jun, 1999; Nadig & Shaw, 2012). Atypical use of affective prosody is one of the core features of ASD. In fact, the emergence of language and speech impairments before three years of age is among the earliest markers of ASD in children (American Psychiatric Association, 2013). A number of descriptions have been offered about the nature of prosodic
impairment in ASD. For example, talkers with ASD have been labeled as sounding from exaggerated to monotonous, and research findings have been contradictory in some cases due to the heterogeneity of the disorder and differences in research methodology (Peppé et al., 2007).

For the estimated 70-80% of individuals with ASD who have developed functional language abilities, atypical prosody is a major contributor to impressions of “social oddness” reported by listeners (Van Bourgondien & Woods, 1992; Paul et al., 2005; Fusaroli et al., 2016), which may hinder the quality of social interaction. However, as discussed by Nadig and Shaw (2012), specific acoustic properties linked to perceptions of oddness have not been identified. Objective acoustic analyses have the potential to enhance assessment of speech production in ASD, and could be used to supplement standard diagnostic instruments such as the Autism Diagnostic Observation Schedule (ADOS, Lord et al., 2000) and Autism Diagnostic Interview (ADI/ADI-R; Le Couteur et al., 1989; Lord, Rutter, & Le Couteur, 1994). Acoustic analysis of affective prosody production offers the ability to assess subtle prosodic differences that may not be detected by behavioral observation alone.

Research involving acoustic analysis of prosody in talkers with ASD has almost exclusively focused on grammatical or pragmatic prosody, such as abnormal sentence stress and intonation patterns, revealing that individuals with ASD exhibit greater f0 range compared to control participants (Fosnot & Jun, 1999; Nadig & Shaw, 2012; Green & Tobin, 2009). In these studies, speech production tasks have involved short phrases produced in interrogative and declarative styles (Fosnot & Jun, 1999), and spontaneous/semi-spontaneous speech produced in English (Nadig & Shaw, 2012) and other languages (Green & Tobin, 2009). The work has revealed important information about prosodic differences in ASD, yet none of the above studies...
specifically involved emotion elicitation. Here, the focus was on determining if previous findings extend to affective prosody, and specifically, to a controlled set of phrases elicited in five emotion contexts: angry, happy, sad, interested, and neutral.

Increased f0 range produced in speaking tasks involving prosody conflicts with traditional accounts of ASD suggesting “flat affect” (e.g., Kanner, 1943) and current diagnostic criteria for ASD which includes “reduced sharing of interests, emotions, or affect” and “a total lack of facial expressions” (American Psychiatric Association, 2013). In speech, flat affect is exhibited by limited modulation in f0 and intensity; reports of increased f0 range in talkers with ASD are inconsistent with flat affect. In the face perception literature, mixed results have been reported regarding reduced expressivity in expressions produced by participants with ASD, and may reflect differences in how the expressions were obtained. Facial expressions obtained from participants with ASD in naturalistic interactions have in some cases shown reduced expressivity compared to controls (Czapinsky & Bryson, 2003; Stagg et al., 2014), whereas recent studies using posed and evoked expressions have reported an opposite pattern – increased expressivity in participants with ASD (Faso, Sasson, & Pinkham, 2015).

One study has examined affective prosody in ASD using emotion elicitation. In Hubbard and Trauner (2007), a set of phrases were elicited in happy, sad, and angry contexts using nine children with ASD, nine children with Asperger’s Syndrome, and ten TD controls. The elicitation methods involved repetition after hearing a recorded actor and a free-response task after hearing a short story. Vowel f0, intensity and duration were extracted from each phrase. In the repetition task, the three talker groups produced expected patterns of emotion-specific acoustic property changes: happy utterances included higher f0 ranges than angry, followed by
sad recordings, however, no interaction was found between group and intended emotion for f0 range. For intensity, systematic differences were found as a function of emotion context: angry phrases contained the greatest intensity, followed by happy and sad productions. While controls used intensity differently for each of the emotions, participants with ASD showed no reliable differences. For duration, the authors found that for control subjects the longest phrase durations were found in the sad phrases, but that talkers with ASD showed no differences between sad phrases and the other emotions.

A key difference between the current study and that of Hubbard and Trauner (2007) is that the current study involved affective prosody recordings using evoked elicitation rather than repetition or free-response. Repetition allows a talker to emulate voice manipulations used by the model talker, and may have led to the absence of talker group effects when separated by emotion context. There is a balance in emotion research between spontaneity and experimental control (Bänziger & Scherer, 2007), and one of the goals for the current database was to examine affective prosody production while controlling for the lexical content of the expressive utterances.

2.2 Method and procedure

2.2.1 Participants

Fifteen adult males with an ASD diagnosis (mean age: 27; age range: 21-42) and fifteen TD males (mean age: 21; age range: 18-26) were recruited as talkers for this study. Participants with ASD were recruited through the University of Texas at Dallas (UTD) Autism Research Collaborative (ARC), which is a database of adults with ASD who expressed interest in research
participation, primarily through a local nonprofit organization. Diagnoses of ASD were confirmed by a certified clinician based on the Autism Diagnostic Observation Schedule (Lord, et al., 2000), and cognitive and language abilities were assessed using the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) and the Wide Range Achievement Test (WRAT-3; Wilkinson, 1993). Participants with ASD were provided with compensation for producing the set of recordings.

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 test for traits associated with the Broad Autism Phenotype (BAP). BAP refers to the presence of autism symptoms that do not reach the categorical boundary for an autism diagnosis (Piven, 2001). Fourteen of fifteen TD talkers had an overall BAPQ score below the self-report cutoff score for males published by Hurley et al. (2007) and the increased specificity score suggested by Sasson et al. (2013). One talker’s BAPQ score surpassed the self-report cutoff score reported by Hurley et al. (2007). The recordings for that talker were analyzed separately and were found to be consistent with other TD talkers. All participants spoke English as a first language and none had a detectable accent. The mean IQ for the ASD group was 107.40 (SD = 12.69), compared to 115.67 (SD = 9.36) for TD talkers. A t-test resulted in no group difference between participants with ASD and controls (p = .053). No group difference was found between talkers with ASD (M=50.27, SD=5.84) and controls (M=49.53, SD=3.27) on the WRAT-3 (p = .675). The study was approved by the UTD
Institutional Review Board, and all participants indicated that they understood the speaking task and provided consent to participate in the research.

2.2.2 Recordings

The recordings were produced in a sound-attenuated booth with the assistance of an experimenter, who led each talker through the recordings protocol using custom Matlab scripts. To minimize experimenter influence on the recordings as much as possible, the protocol included pre-recorded audio instructions presented through computer speakers with matching text instructions displayed on a computer screen. The recording equipment included a Shure SM-94 microphone, Symetrix SX202 pre-amplifier and Tucker-Davis Technologies System III data acquisition hardware (MA1, RP2.1). Digital waveforms were stored on computer disk at a rate of 48 kHz and 16-bit resolution. A previous experiment using a similar elicitation technique and identical recording conditions is described in Hubbard and Assmann (2013; Experiment II).

Each recording was produced in the following five emotion contexts: neutral, angry, happy, interested, and sad. The emotion contexts were selected because they occupy distinct regions along two dimensions known to be important for human emotional experience: valence and arousal level (see Figure 1.1). Valence and arousal level have long been considered to be fundamental aspects of emotion, dating back to the pioneering work of Wilhelm Wundt (1909) and many emotion researchers since that time (e.g., Schlosberg, 1952, 1954; Russell, 1980; Larsen & Diener, 1992). Happy and angry contexts occupy positive and negative ends of the high arousal dimension, and the interested and sad contexts occupy positive and negative ends of the low arousal dimension (Russell, 2003). Recordings produced in a neutral context were collected to provide a baseline for comparison with the other contexts.
Evoked and portrayed emotion elicitation methods were used to obtain the expressive recordings. In the evoked method, talkers were asked to recall personal experiences that made them feel angry, happy, interested, or sad and remember details of those situations prior to producing the following set of five phrases: 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?” For the neutral context, talkers were asked to recall a time when they didn’t feel any particular emotion at all prior to producing the phrases. The evoked elicitation method has been used effectively in previous work, particularly in studies involving face perception (e.g., Faso, Sasson, & Pinkham, 2015).

The phrases were chosen because of their ambiguity – each can be realistically produced in any of the five emotion contexts. This is an important methodological aspect of the project that allowed group-based comparisons of affective prosody production using the same verbal content. Each talker was given as much time as needed to recall an emotional episode and to orient themselves to the appropriate mindset for each expressive production. With the exception of the neutral context, talkers were asked to speak each phrase three times in increasingly expressive repetitions (e.g., happy, happier, happiest) to assess the ability to enhance expressiveness from the first to last production in autistic and non-autistic talkers. The experimenter provided each talker with a printed copy of the ambiguous phrases to refer to during the evoked elicitation procedure, and pre-recorded instructions were presented to the talkers over speakers and displayed as text on the computer screen.

In the portrayed elicitation method, each talker was presented with an emotion prompt designed to help familiarize them with the desired context. The emotion prompts were presented
through computer speakers, and text versions of the prompts were displayed on a computer screen. As in the evoked phrases, the instructions and prompts were embedded in the recordings protocol implemented using custom Matlab scripts. Prompts consisted of the following topics: lost luggage (angry), winning the lottery (happy), a new job opportunity (interested), a terminal illness diagnosis (sad), and a superficial conversation with a fellow student after class (neutral).

Scripted dialogue sessions were recorded first in which the talkers interacted with the computer using pre-recorded dialogue and target phrases displayed on computer screen. As an example, the dialogue session for the happy context is detailed in Table 2.1. The dialogue sessions for all five emotion contexts consisted of three short pre-recorded phrases and three target phrases produced by each talker. During the dialogue sessions talkers were asked to listen carefully to the emotion prompts and imagine each situation is happening to them, then engage in the dialogue session by producing the target phrases (displayed in green text on the screen) in the prompted emotion context.

Table 2.1. Example dialogue script for the happy emotion context

| Scripted Dialogue: Happy Expressive prompt: Winning the lottery | Pre-recorded 1: “What happened” | Target 1: “I can’t believe this!” | Pre-recorded 2: “Can’t believe what?” | Target 2: “I just won the lottery!” | Pre-recorded 3: “No way, are you serious?” | Target 3: “Yes…yes…yes! I matched all six numbers!” |

After each scripted dialogue session, the talkers produced a set of eight nonsense vowel-consonant-vowel (VCV) syllables: aba, aga, ibi, igi, obo, ogo, ubu, and ugu. The VCV recordings were collected to allow comparison of affective prosody in utterances in which semantic content is absent. In the scripted dialogues talkers were able to match their affective tone with the lexical content in the target phrase (e.g., “I just won the lottery” is congruent with
and easy to produce in a happy voice, but perhaps not as easy in a sad voice). In the evoked phrases and nonsense VCVs, however, the lexical content was neutral, thus, clues to indicate an emotion context were absent. The VCV syllables were collected to investigate affective prosody production in speech in which semantic content was absent.

The VCV syllables were produced twice – once before and once after the talkers listened to an example set of recordings produced by the experimenter. This was done so that the acoustic properties of pre- and post-example recordings could be compared to assess whether talkers with ASD respond in the same way as TD talkers to hearing the example. The recordings protocol for the portrayed elicitation method was completed in the following order per emotion context. First, the emotion prompt was presented audibly and in text on the computer screen. Second, the scripted dialogue session was completed, followed by the isolated vowel “ah” and the VCV syllables (see Figure 2.1).

As described above, for each of the five emotion contexts, the fifteen talkers with ASD and fifteen TD talkers produced the five ambiguous phrases three times using the evoked elicitation method. This resulted in 75 evoked phrases (15 utterances in five emotion contexts). Using the prompted elicitation method, each talker produced the scripted dialogue session consisting of three utterances, and the eight VCV syllables in each emotion context. The prompted elicitation method yielded 130 utterances per talker, and the two elicitation methods in sum resulted in 205 recordings per talker. ASD and TD talkers consistently completed the recordings protocol in just over one hour.
The obtained expressive speech recordings extend the study of affective prosody communication in ASD in several ways. A key aspect of the database is that it included recordings in which the same lexical context is produced by each talker in multiple emotion contexts. This is in contrast with studies in which spontaneous or semi-spontaneous speech is used (Nadig & Shaw, 2012), and those in which different verbal content was used for different emotions (Hubbard & Trauner, 2007). The current methodology offered the ability to make broad group comparisons as in previous studies, but also allowed more detailed analyses such as
comparisons between the neutral context and other emotions, differences in segments of speech within phrases, and within-group comparisons using speech with identical verbal content.

An additional way the current database extends prior studies is that it included nonsense vowel-consonant-vowel (VCV) syllables and phrases that are ambiguous in their semantic content. In contrast with the evoked phrases containing ambiguous lexical content, and the scripted dialogue sessions for which the linguistic content is consistent with the target emotion, the lexical content of the VCV syllables and isolated vowels was inaccessible to talkers to signal the expected emotional content to be portrayed. Few collections of ASD and TD expressive speech recordings exist in which the two groups of talkers produced the same speech materials in different emotion contexts. Thus, the design allowed detailed acoustic analyses to compare ASD and TD talkers using a controlled speech corpus containing multiple emotion contexts and speech types.
CHAPTER 3
ACOUSTIC ANALYSIS OF AFFECTIVE PROSODY

3.1 Introduction

This chapter presents the results of acoustic analyses performed on the expressive speech recordings described in Chapter 2. The purpose of the acoustic analysis was to identify voice properties that are relevant for communication of affective prosody, and to determine whether production of those properties differs for talkers with ASD compared to TD controls. Research using TD participants has shown that emotions produced in speech can be identified based on differences in the production patterns of f0, intensity, and duration (Scherer, 1979, 1986; Murray & Arnott, 1993), and that those differences are known to be important for listener perception (Tóth, Sztahó, & Vicsi, 2008; Waaramaa et al., 2010).

In participants with ASD, research has revealed divergent patters of affective prosody production compared to TD talkers, but the studies have primarily involved grammatical or pragmatic prosody (Simmons & Baltaxe, 1975; Fosnot & Jun, 1999; Nadig & Shaw, 2012) as opposed to affective prosody obtained by direct emotion elicitation (but see Hubbard & Trauner, 2007). The body of work published to date has shown that talkers with ASD produce emotional content in conversational speech with increased f0 range compared to TD talkers. The acoustic analysis presented in this chapter examined whether increased f0 range is found using direct emotion elicitation and a controlled set of speech materials collected from adult males with ASD.
Very little research has investigated systematic differences in intensity and speaking duration in talkers with ASD. In Hubbard & Trauner (2007), differences in amplitude range were found in angry, happy, and sad speech. The highest mean amplitude levels were found in angry speech, and the lowest was found in sad speech. No interaction was found, however, when the talker groups were compared in each emotion context. The ASD group did not differ significantly from the control group in intensity between the emotions, but the control group produced a higher intensity range in the angry context and no difference was found in the happy and sad utterances.

The original database recordings were processed using software designed to provide estimates for a range of acoustic properties across the duration of each utterance. Acoustic properties related to speech production can be divided in two basic categories – voice source and vocal tract filter properties. Voice source properties are those related to the source of voicing—the larynx and vibration of the vocal folds, whereas vocal tract filter properties are related to the shape of the oral and nasal airways that modulate the passage of sound energy from the voicing source. The acoustic analysis included in this chapter focused on voice source properties, given that the link between the voice source and production of affective prosody has been established for some time (Lieberman & Michaels, 1962; Murray & Arnott, 1993; Scherer, 2003).

Much of the previous work in this area has involved summary measures of f0 such as the mean and range of the f0 contour; however, detailed analysis of the f0 contour was included here through the use of functional data analysis (Ramsay & Silverman, 2005), which is a collection of statistical techniques designed for analyzing temporal trajectories that naturally exist as curves or functions (such as f0 contours). The acoustic measurements were analyzed to reveal patterns of
affective prosody production specific to each elicited emotion context and differences between the two talker groups.

3.2 Method and procedure

The recordings were pre-processed using a semi-automated procedure implemented with custom Matlab scripts. The procedure isolates individual utterances (when more than one utterance was obtained per recording), truncates silence at the beginning and end of each phrase, and saves the truncated files separately for acoustic analysis. The recordings were then processed using a set of algorithms designed to measure acoustic properties at different time points along the duration of each utterance. The algorithms operate by dividing each recording into 1 ms analysis frames and estimating the set of acoustic parameters across multiple frames. The following section describes the set of acoustic parameters obtained and the method used to estimate each parameter.

Fundamental frequency (f0, perceived as voice pitch) was estimated using STRAIGHT (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999), a high-quality analysis and synthesis tool that decomposes input speech into separate elements related to the voicing source and vocal tract filter. F0 was included in the earliest systematic investigations of emotion production in speech (Fairbanks & Provonost, 1939; Williams & Stevens, 1972), and is known to be a dominant voice cue important for perception of a talker’s affective state. One reason for this is that talkers are relatively free to use changes in f0 to convey non-linguistic information in speech, whereas the vocal tract filter plays a primary role in conveying linguistic content (Williams & Stevens, 1972). F0 was estimated across multiple analysis frames and measurements were output at each analysis frame along the duration of each recording. The f0 data were time-normalized for each of the
three speech types by interpolating the f0 estimates using a fixed number of data points – evoked phrases were represented by 500 analysis points, VCV syllables were normalized to 200 points, and scripted dialogue sessions were normalized to 1000 points. Segments of silence between words and before and after utterances were removed, and the number of data points corresponded to approximately half of the average duration in ms for each speech type.

Intensity (dB SPL) estimates were obtained using Praat (Boersma & Weenink, 2014), an open source speech analysis software package. Praat scripts were developed to automatically extract intensity estimates at each analysis frame corresponding to 1 ms slices through the duration of each evoked phrase, VCV, and scripted dialogue phrase. 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, including unvoiced segments between words. Unvoiced segments were included in duration measurements because patterns of silence between words in a phrase may have relevance for perceptions of social oddness when communicating with a talker with ASD.

The acoustic analysis presented in this chapter is divided into two sections, consisting of ANOVAs using summary statistics based on the above acoustic measures, and more detailed analysis of the f0 and intensity contours using functional data analysis (Ramsay & Silverman, 2005). Both methods were used to identify acoustic 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. For the AVOVAs, summary measures of each acoustic property including the range of f0, and mean calculations for intensity and duration were computed and subjected to analysis. 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. Additional group comparisons were performed on the scripted dialogue recordings to determine if any identified patterns of divergent affective prosody production carried over to the dialogue sessions. The comparison of results between evoked phrases and prompted VCV syllables and dialogues was designed to provide insight into production differences as a function of elicitation method, speech type, and lexical ambiguity.

3.3 Hypotheses

Voice recordings collected from talkers with ASD were expected to differ significantly from those collected from TD talkers. It was predicted that acoustic properties relevant to affective prosody produced by talkers with ASD would show more variable patterns of production compared to those produced by TD talkers. This prediction was based on the results of previous studies that reported systematic differences in recordings obtained from talkers with ASD compared to controls (Simmons & Baltaxe, 1975; Shriberg et al., 2001), including increased f0 variability, as measured by the range of f0 across a given utterance (Fosnot & Jun, 1999; Diehl et al., 2009; Nadig & Shaw, 2012).

Due to differences in processing of affective prosody and the resulting impact on learned patterns of prosody production, ASD talkers were expected to exhibit inconsistent modulation patterns in emotion-related acoustic properties compared to controls. Consistent with prior work, it was predicted that both groups would show similar types of acoustic changes during productions of different emotion contexts, but that those produced by talkers with ASD would be less consistently produced.
Production variability was measured in several ways. The standard deviation of f0 (in Hz), intensity (in dB SPL), and duration (in ms) was calculated for each utterance and summarized for each group. Additionally, the f0 and intensity contours of each utterance were plotted with 95% confidence bands superimposed to provide another measure of production variability. A significantly higher standard deviation of f0, intensity, and duration, and wider confidence bands surrounding the f0 and intensity contours for the ASD group would suggest less consistent patterns of production for the ASD talkers.

3.4 Innovation

The acoustic analysis presented in this chapter investigated differences between talkers with ASD and TD controls in the three most commonly reported acoustic parameters relevant for affective prosody production (f0, intensity, and duration) using evoked and portrayed emotion elicitation. Compared with prior research, the analysis contained a more comprehensive examination of dynamic modulations that accompany production of affective prosody. Global measures such as the mean and range of f0 provide important information about general trends in emotion-related acoustic properties; however, a more detailed assessment of the specific ways in which relevant properties modulate in the expressive speech of talkers with ASD and TD controls is presented using functional data analysis (Ramsay & Silverman, 1997). For example, production differences involving changes to constantly varying acoustic properties may not emerge in comparisons using summary measures alone. Functional data analysis offered a fine-grained analysis of the f0 and intensity patterns produced by TD talkers, and to what extent they deviate in emotion portrayals produced by ASD talkers. The technique has been used to identify
reliable markers of different emotion contexts in typically-developing talkers (Arias, Busso, & Yoma, 2014).

3.5 Results

The discussion of the acoustic analysis results is divided according to the type of speech produced, and consists of subsections for the evoked ambiguous phrases (see Section 3.5.1), prompted VCVs (see Section 3.5.2) and scripted dialogue sessions (see Section 3.5.3).

3.5.1 Evoked phrases

Each talker produced 75 evoked phrases, consisting of three repetitions of five phrases produced in five emotion contexts. 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 3.1 lists the mean, standard deviation, and range for f0, intensity, and duration for evoked phrases for each emotion context and talker group.

<table>
<thead>
<tr>
<th>context</th>
<th>group</th>
<th>N</th>
<th>f0 (Hz) mean</th>
<th>range</th>
<th>Intensity (dB SPL) mean</th>
<th>range</th>
<th>Duration (ms) mean</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutral</td>
<td>ASD</td>
<td>225</td>
<td>119.0</td>
<td>46.4</td>
<td>65.4</td>
<td>3.7</td>
<td>1154.7</td>
<td>273.6</td>
</tr>
<tr>
<td>neutral</td>
<td>TD</td>
<td>225</td>
<td>108.3</td>
<td>61.7</td>
<td>63.6</td>
<td>3.0</td>
<td>1147.7</td>
<td>277.1</td>
</tr>
<tr>
<td>angry</td>
<td>ASD</td>
<td>225</td>
<td>176.4</td>
<td>86.9</td>
<td>67.3</td>
<td>4.2</td>
<td>1173.8</td>
<td>303.6</td>
</tr>
<tr>
<td>angry</td>
<td>TD</td>
<td>225</td>
<td>128.0</td>
<td>56.4</td>
<td>66.0</td>
<td>3.3</td>
<td>1150.1</td>
<td>254.0</td>
</tr>
<tr>
<td>happy</td>
<td>ASD</td>
<td>225</td>
<td>198.8</td>
<td>106.7</td>
<td>67.5</td>
<td>3.4</td>
<td>1187.2</td>
<td>349.3</td>
</tr>
<tr>
<td>happy</td>
<td>TD</td>
<td>225</td>
<td>159.4</td>
<td>74.8</td>
<td>66.4</td>
<td>3.4</td>
<td>1035.1</td>
<td>228.9</td>
</tr>
<tr>
<td>interested</td>
<td>ASD</td>
<td>225</td>
<td>158.0</td>
<td>89.3</td>
<td>65.4</td>
<td>3.5</td>
<td>1323.0</td>
<td>372.1</td>
</tr>
<tr>
<td>interested</td>
<td>TD</td>
<td>225</td>
<td>134.7</td>
<td>62.2</td>
<td>65.4</td>
<td>4.2</td>
<td>1063.4</td>
<td>264.3</td>
</tr>
<tr>
<td>sad</td>
<td>ASD</td>
<td>225</td>
<td>150.6</td>
<td>73.9</td>
<td>65.7</td>
<td>4.4</td>
<td>1372.2</td>
<td>398.2</td>
</tr>
<tr>
<td>sad</td>
<td>TD</td>
<td>225</td>
<td>122.2</td>
<td>61.8</td>
<td>63.2</td>
<td>3.3</td>
<td>1228.6</td>
<td>341.9</td>
</tr>
</tbody>
</table>

Table 3.1. Evoked phrase acoustic data: the number of phrases (N), group means, standard deviations, and ranges for f0, intensity, and duration by emotion context and talker group.
An ANOVA was performed on each of the acoustic measures, treated as dependent variables, and group (ASD vs. TD), context (neutral, angry, happy, interested, & sad), phrase (five ambiguous phrases) and repetition (three repetitions) were treated as independent variables. Talkers were treated as a repeated random variable given that the same group of talkers produced multiple recordings included in the analysis.

Figure 3.1. 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.

The analysis of f0 range revealed main effects for talker group ($F (1, 28) = 4.648, p < .05$), emotion context ($F (4, 2069) = 110.504, p < .001$) and phrase ($F (4, 2069) = 13.548, p < .001$), and interactions between talker group and emotion context ($F (4, 2069) = 14.290, p < .001$), talker group and phrase ($F (4, 2069) = 2.451, p < .05$), and emotion context by phrase ($F (16, 2069) = 3.947, p < .001$). 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 3.1 shows mean f0 range calculations by group and context. Phrases produced by talkers with ASD had a mean f0 range of 119 Hz ($SD = 91 \text{ Hz}$), compared to 93 Hz ($SD = 67 \text{ Hz}$) for TD talkers.

![Figure 3.1](image1.png)

**Figure 3.1.** 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.

Individual f0 contours for the phrase “Why did you do that” are shown in Figure 3.2 (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.

Figure 3.3. 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.

The analysis of mean intensity resulted in main effects for talker group ($F(1, 28) = 4.221, p < .05$), emotion context ($F(4, 2072) = 69.798, p < .001$) and phrase ($F(4, 2072) = 49.834, p < .001$). Significant interactions were found for talker group by emotion context ($F(4, 2072) = 10.723, p < .001$), talker group by phrase ($F(4, 2072) = 5.354, p < .001$), emotion context by phrase ($F(16, 2072) = 3.305, p < .001$), and emotion context by repetition ($F(8, 2072) = 8.908, p < .001$). The three-way interaction between talker group, emotion context and phrase was also significant ($F(16, 2072) = 2.213, p < .05$). 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 performed for each emotion context (repetitions collapsed) 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 3.3 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. Table 3.1 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 ($F (4, 2057) = 46.263, p < .001$), phrase ($F (4, 2057) = 218.891, p < .001$), and repetition ($F (4, 2057) = 10.243, p < .001$). There was no significant group difference in mean duration for the evoked phrases, although the result approached significance ($p = .069$). There were significant interactions between talker group and emotion context ($F (4, 2057) = 24.232, p < .001$), talker group and phrase ($F (4, 2057) = 10.263, p < .001$), and emotion context and phrase ($F (16, 2057) = 3.345, p < .001$). 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 3.4 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 3.4](image)

Figure 3.4. 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.

### 3.5.2 Prompted VCV syllables

Each talker provided recordings of 80 VCV syllables, consisting of eight syllables in five emotion contexts produced in two conditions: the first prior to hearing a pre-recorded example spoken by an experimenter, and the second after. For each utterance, f0 range (Hz), mean intensity (dB SPL), and duration (ms) were obtained to examine acoustic differences in affective prosody production between the ASD and TD talker groups. An ANOVA was performed on
each of the acoustic measures, treated as dependent variables, and talker group (ASD vs. TD), emotion context (neutral, angry, happy, interested, & sad), syllable (eight VCVs) and condition (pre- or post-example) were treated as independent variables. Talkers were treated as a repeated random variable given that the same group of talkers produced multiple recordings included in the analysis. Table 3.2 lists the mean, standard deviation, and range for f0, intensity, and duration in VCV syllables for each emotion context and talker group.

Table 3.2. VCV acoustic data: number of syllables (N), group mean, standard deviation, and range for f0, intensity, and duration by emotion context and talker group.

<table>
<thead>
<tr>
<th>context</th>
<th>group</th>
<th>N</th>
<th>f0 (Hz) mean</th>
<th>SD</th>
<th>range</th>
<th>Intensity (dB SPL) mean</th>
<th>SD</th>
<th>range</th>
<th>Duration (ms) mean</th>
<th>SD</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutral</td>
<td>ASD</td>
<td>240</td>
<td>111.4</td>
<td>23.8</td>
<td>31.8</td>
<td>67.4</td>
<td>3.6</td>
<td>21.1</td>
<td>483.8</td>
<td>74.1</td>
<td>575.0</td>
</tr>
<tr>
<td>neutral</td>
<td>TD</td>
<td>234</td>
<td>102.8</td>
<td>29.6</td>
<td>28.6</td>
<td>65.7</td>
<td>3.2</td>
<td>16.2</td>
<td>477.7</td>
<td>83.2</td>
<td>413.0</td>
</tr>
<tr>
<td>angry</td>
<td>ASD</td>
<td>240</td>
<td>134.1</td>
<td>39.9</td>
<td>53.2</td>
<td>69.2</td>
<td>4.1</td>
<td>22.2</td>
<td>491.9</td>
<td>103.9</td>
<td>487.0</td>
</tr>
<tr>
<td>angry</td>
<td>TD</td>
<td>237</td>
<td>109.6</td>
<td>30.1</td>
<td>44.0</td>
<td>66.8</td>
<td>4.3</td>
<td>19.3</td>
<td>476.0</td>
<td>66.2</td>
<td>344.0</td>
</tr>
<tr>
<td>happy</td>
<td>ASD</td>
<td>240</td>
<td>146.0</td>
<td>60.0</td>
<td>76.5</td>
<td>68.7</td>
<td>3.9</td>
<td>20.2</td>
<td>510.1</td>
<td>103.2</td>
<td>540.0</td>
</tr>
<tr>
<td>happy</td>
<td>TD</td>
<td>236</td>
<td>133.8</td>
<td>46.2</td>
<td>79.7</td>
<td>67.1</td>
<td>3.6</td>
<td>20.6</td>
<td>473.5</td>
<td>69.6</td>
<td>338.0</td>
</tr>
<tr>
<td>interested</td>
<td>ASD</td>
<td>240</td>
<td>141.1</td>
<td>53.1</td>
<td>75.5</td>
<td>67.7</td>
<td>3.7</td>
<td>18.5</td>
<td>532.6</td>
<td>100.0</td>
<td>510.0</td>
</tr>
<tr>
<td>interested</td>
<td>TD</td>
<td>237</td>
<td>118.4</td>
<td>37.6</td>
<td>60.8</td>
<td>65.7</td>
<td>3.4</td>
<td>18.3</td>
<td>484.0</td>
<td>88.6</td>
<td>430.0</td>
</tr>
<tr>
<td>sad</td>
<td>ASD</td>
<td>240</td>
<td>120.4</td>
<td>44.2</td>
<td>45.7</td>
<td>67.1</td>
<td>3.8</td>
<td>22.9</td>
<td>568.3</td>
<td>109.8</td>
<td>706.0</td>
</tr>
<tr>
<td>sad</td>
<td>TD</td>
<td>238</td>
<td>102.7</td>
<td>26.3</td>
<td>34.3</td>
<td>63.5</td>
<td>3.9</td>
<td>24.9</td>
<td>555.8</td>
<td>103.1</td>
<td>539.0</td>
</tr>
</tbody>
</table>

For f0 range, main effects were found for emotion context ($F (4, 2178) = 122.953, p < .001$) and syllable ($F (7, 2178) = 6.457, p < .001$). The talker group by emotion context ($F (4, 2178) = 4.240, p = .002$) and the emotion context by condition ($F (4, 2178) = 3.228, p = .012$) interactions were also significant. Condition (pre- vs. post-example productions) and the interaction between talker group and condition were not significant, which shows that, in general, ASD and TD talkers did not significantly alter their patterns of f0 production after hearing the experimenter example. The mean f0 range for the ASD talker group was 56.40 Hz ($SD = 48.86$ Hz) compared to 49.47 Hz ($SD = 39.29$) for the TD talker group. Planned comparisons for each emotion context revealed that f0 variability, as defined by the mean f0
range for each utterance, was greater for talkers with ASD in the interested and sad emotion context, and approached significance in the angry condition ($p = .066$). Figure 3.5 shows mean f0 range calculations for VCV syllables by talker group and emotion context.

![Figure 3.5. Mean f0 range (in Hz) for VCV syllables produced by talkers with ASD and TD controls. Bars represent talker group f0 range means per emotion context. Error bars represent +/- one mean standard error.](image)

The analysis of intensity resulted in main effects for talker group ($F (1, 28) = 15.830, p < .001$), emotion context ($F (4, 2194) = 61.345, p < .001$), syllable ($F (7, 2194) = 53.194, p < .001$), and condition ($F (1, 2194) = 5.863, p = .015$). Significant interactions were found for talker group by emotion context ($F (4, 2194) = 7.147, p < .001$), emotion context by condition ($F (4, 2194) = 4.040, p = .003$), and the three-way interaction between talker group, emotion context and condition was significant ($F (4, 2194) = 4.459, p = .001$). Bonferroni-corrected post-hoc tests for each emotion context revealed that syllables produced by talkers with ASD in each emotion
context were spoken with greater mean intensity than those produced by TD talkers. The mean intensity for talkers with ASD was 68.03 dB SPL ($SD = 3.93$ dB SPL), compared to 65.73 dB SPL ($SD = 3.70$ dB SPL) for the control group. Figure 3.6 shows mean intensity by talker group, emotion context and condition. The mean intensity for the experimenter examples is represented for each emotion context with a black triangle.

![Figure 3.6](image)

Figure 3.6. Mean intensity levels (dB SPL) for VCV syllables produced by talkers with ASD and TD controls. Points represent means for each talker group, emotion context and condition (before and after a recorded example produced by the experimenter was presented). Error bars represent $+/-$ one mean standard error. The means for the experimenter examples are represented by black triangles.

Intensity was also compared at 10% slices along the duration of each syllable. Figure 3.7 displays mean intensity contours with group comparisons at each of the 10% slices for the angry, happy, interested and sad contexts. Figure 3.7 shows the main effect of talker group on mean
intensity. Talkers with ASD produced the syllables with greater intensity compared to TD talkers; however, this occurred only in the stressed portion of the syllables. For the middle portion of analysis points where the glottal stop for the consonant /b/ or /g/ occurred, no group differences were found, or the opposite pattern was found showing higher intensity for TD talkers.

Figure 3.7. Mean intensity contours for VCV syllables produced by talkers with ASD and TD controls. Points represent talker group means calculated at each 10% slice through the time-normalized duration of the syllables. Error bars represent +/- one mean standard error calculated per slice.

For duration, main effects were found for emotion context ($F(4, 2177) = 115.542, p < .001$), syllable ($F(7, 2177) = 9.134, p < .001$), and condition ($F(1, 2177) = 39.943, p < .001$), and the interaction between talker group by emotion context was significant ($F(4, 2177) = 8.836, p < .01$). Bonferroni-corrected post-hoc tests were conducted to compare talker groups for
each emotion context and revealed that talkers with ASD produced longer syllables in the happy and interested contexts. Figure 3.8 shows mean syllable durations by group, context, and condition. Mean syllable duration for talkers with ASD was 517 ms ($SD = 103$ ms) compared to 464 ms ($SD = 89$ ms) for TD talkers.

![Graph showing mean syllable duration for VCV syllables produced by talkers with ASD and TD controls.](image)

Figure 3.8. Mean duration (in ms) for VCV syllables produced by talkers with ASD and TD controls. Bars represent means for each talker group, emotion context and condition (before and after a recorded example produced by the experimenter was presented). Error bars represent +/- one mean standard error. The mean durations of the experimenter examples are represented by black triangles.

### 3.5.3 Scripted dialogue sessions

Each talker produced 15 phrases embedded within the scripted dialogue sessions, consisting of three utterances in five emotion contexts. For each utterance, $f_0$ range (Hz), mean intensity (dB SPL), and duration (ms) were calculated to examine acoustic differences in affective prosody production between the ASD and TD groups. An ANOVA was performed on
each of the acoustic measures, treated as dependent variables, and group (ASD vs. TD) and context (neutral, angry, happy, interested, & sad) were treated as independent variables. Talkers were treated as a repeated random variable given that the same group of talkers produced multiple recordings included in the analysis. Table 3.3 lists the mean, standard deviation, and range for f0, intensity, and duration in scripted dialogues for each emotion context and talker group.

Table 3.3. Scripted dialogue acoustic data: number of phrases (N), group mean, standard deviation, and range for f0, intensity, and duration by emotion context and talker group.

<table>
<thead>
<tr>
<th>context</th>
<th>group</th>
<th>N</th>
<th>f0 (Hz) mean</th>
<th>f0 (Hz) SD</th>
<th>f0 (Hz) range</th>
<th>Intensity (dB SPL) mean</th>
<th>Intensity (dB SPL) SD</th>
<th>Intensity (dB SPL) range</th>
<th>Duration (ms) mean</th>
<th>Duration (ms) SD</th>
<th>Duration (ms) range</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutral</td>
<td>ASD</td>
<td>45</td>
<td>118.5</td>
<td>31.3</td>
<td>47.9</td>
<td>61.6</td>
<td>3.3</td>
<td>14.9</td>
<td>1980.8</td>
<td>908.1</td>
<td>3134.0</td>
</tr>
<tr>
<td>neutral</td>
<td>TD</td>
<td>45</td>
<td>108.0</td>
<td>25.6</td>
<td>38.8</td>
<td>62.0</td>
<td>3.1</td>
<td>11.9</td>
<td>1813.0</td>
<td>798.4</td>
<td>2650.0</td>
</tr>
<tr>
<td>angry</td>
<td>ASD</td>
<td>45</td>
<td>194.1</td>
<td>111.6</td>
<td>161.1</td>
<td>67.6</td>
<td>3.1</td>
<td>12.8</td>
<td>3049.6</td>
<td>1324.5</td>
<td>5482.0</td>
</tr>
<tr>
<td>angry</td>
<td>TD</td>
<td>45</td>
<td>132.0</td>
<td>50.8</td>
<td>81.7</td>
<td>65.3</td>
<td>2.6</td>
<td>12.0</td>
<td>2898.8</td>
<td>1172.7</td>
<td>3513.0</td>
</tr>
<tr>
<td>happy</td>
<td>ASD</td>
<td>45</td>
<td>226.7</td>
<td>93.4</td>
<td>164.8</td>
<td>68.1</td>
<td>3.4</td>
<td>15.2</td>
<td>2300.2</td>
<td>1622.9</td>
<td>5193.0</td>
</tr>
<tr>
<td>happy</td>
<td>TD</td>
<td>45</td>
<td>166.1</td>
<td>73.0</td>
<td>117.8</td>
<td>64.9</td>
<td>4.0</td>
<td>15.7</td>
<td>2163.0</td>
<td>1768.9</td>
<td>5943.0</td>
</tr>
<tr>
<td>interested</td>
<td>ASD</td>
<td>45</td>
<td>147.9</td>
<td>68.5</td>
<td>99.7</td>
<td>63.4</td>
<td>2.8</td>
<td>12.1</td>
<td>2248.5</td>
<td>315.0</td>
<td>1303.0</td>
</tr>
<tr>
<td>interested</td>
<td>TD</td>
<td>45</td>
<td>123.8</td>
<td>28.4</td>
<td>60.9</td>
<td>63.4</td>
<td>2.8</td>
<td>14.8</td>
<td>2132.5</td>
<td>305.1</td>
<td>1305.0</td>
</tr>
<tr>
<td>sad</td>
<td>ASD</td>
<td>45</td>
<td>137.1</td>
<td>52.7</td>
<td>60.9</td>
<td>62.8</td>
<td>2.8</td>
<td>11.9</td>
<td>3057.4</td>
<td>1384.9</td>
<td>4680.0</td>
</tr>
<tr>
<td>sad</td>
<td>TD</td>
<td>45</td>
<td>113.0</td>
<td>29.2</td>
<td>46.4</td>
<td>61.0</td>
<td>2.8</td>
<td>13.7</td>
<td>2723.6</td>
<td>1207.9</td>
<td>4015.0</td>
</tr>
</tbody>
</table>

For f0 range, main effects were found for talker group ($F(1, 28) = 15.577, p < .001$), emotion context ($F(4, 391) = 59.171, p < .001$), and phrase ($F(2, 391) = 27.640, p < .001$). The interactions between talker group and emotion context ($F(4, 391) = 6.556, p < .001$) and emotion context and phrase ($F(8, 391) = 9.562, p < .001$) were also significant. Differences in f0 range by emotion context and phrase were expected, given that unique patterns of f0 variation are known to accompany emotional speech in separate emotion contexts, and that each phrase contained different lexical content. The mean f0 range for scripted dialogue phrases produced by talkers with ASD was 106.90 Hz ($SD = 90.75$ Hz), compared to 69.10 Hz ($SD = 53.09$) for TD talkers. See Table 3.3 for more a more detailed acoustic summary by emotion context. Planned
comparisons to assess talker group differences within each context confirmed the prediction that f0 variability (as defined by mean f0 range) was greater for each emotion context except neutral in phrases produced by talkers with ASD. Figure 3.9 displays mean f0 range calculations by talker group, emotion context and phrase number.

Figure 3.9. Mean f0 range (in Hz) for scripted dialogue recordings produced by talkers with ASD and TD controls. Bars represent means for each talker group and emotion context. Error bars represent +/- one mean standard error.

The analysis of mean intensity (dB SPL) revealed main effects for talker group ($F(1, 28) = 5.432, p = .027$), emotion context ($F(4, 392) = 78.946, p < .001$) and phrase ($F(2, 392) = 9.047, p < .001$). The interactions between talker group by emotion context ($F(4, 392) = 8.424, p < .001$), emotion context by phrase ($F(8, 392) = 8.408, p < .001$), and the three-way interaction between talker group, emotion context and phrase were also significant ($F(8, 392) = 2.006, p = .045$). Figure 3.10 displays mean intensity by talker group, emotion context and phrase number.
for the scripted dialogue recordings. Overall, the mean intensity for the ASD talker group was 64.70 dB SPL \((SD = 4.06 \text{ dB SPL})\) compared to 63.33 dB SPL \((SD = 3.46 \text{ dB SPL})\) for the TD talker group.

![Figure 3.10: Mean intensity (in dB SPL) for scripted dialogue recordings produced by talkers with ASD and TD controls. Points represent means for each talker group, emotion context and phrase number. Error bars represent +/- one mean standard error.](image)

The analysis of phrase duration revealed main effects for emotion context \((F(4, 398) = 134.58, p < .001)\) and phrase \((F(2, 398) = 1037.80, p < .001)\), but the difference in duration between talker groups did not reach significance \((p = .068)\). The interaction between emotion context and phrase number was significant \((F(8, 398) = 181.28, p < .001)\). Figure 3.11 displays mean phrase durations by group, context, and condition. The mean duration for scripted dialogue phrases produced by talkers with ASD was 2528.54 ms \((SD = 1274.33 \text{ ms})\) compared to 2346.20 ms \((SD = 1215.69 \text{ ms})\) for TD talkers.
3.6 Functional data analysis

3.6.1 Introduction

Functional data analysis (FDA) is a set of statistical methods used for analyzing data that naturally exist as curves or functions (Ramsay & Silverman, 1997). Acoustic data such as f0 and intensity contours are ideally suited for FDA, given the temporally-varying and continuous nature of the acoustic measurements. FDA was performed on the evoked phrases to further investigate modes of variation in production of affective prosody among talkers with ASD and TD controls. The following sections describe the FDA analysis procedure and results, which complement the results presented in section 3.5.
An advantage of FDA is that rather than using descriptive statistics to summarize temporally-varying acoustic data, it is possible to analyze entire functions. In contrast to using a single measure to describe a dynamic acoustic property such as f0 (e.g., f0 range), FDA provides more detailed information about the shape of the contour and features that explain large portions of variance in the functional data. For example, FDA makes it possible to determine not only global differences in two or more productions of the same speech content in the same emotion context but also where along f0 and intensity contours the utterances are reliably different.

FDA has been used to develop statistical models for emotion detection (Arias, Busso, & Yoma, 2014) and to examine patterns of intonation (Zellers, Gubian, & Post, 2010) and vowel reduction in conversational speech (Gubian, et al., 2009) in TD talkers. In the current application, the functional extensions to principal component analysis (PCA) and ANOVA were used to supplement the descriptive statistics presented in section 3.5. Functional PCA was used as a data reduction method to identify key sources of variability in the functional data, and functional ANOVA was used to test for significant differences between groups of talkers with ASD and typically-developing controls.

3.6.2 Method and procedure

Traditional multivariate PCA was developed as a method of converting data sets of inter-correlated multidimensional variables into uncorrelated variables (Pearson, 1901; Hotelling, 1933). The goals of PCA are to extract and retain only the most important information from a data set and to analyze the structure of the observations which are typically represented by rows in the data set, and variables, which are typically represented by columns (Abdi & Williams, 2010). PCA computes new variables called principal components which are orthogonal linear
combinations of the original variables. The relevance of the principal components is then reflected by the amount of total variance explained by each component. In functional PCA, a set of functions representing f0 or intensity contours are statistically represented to extract principal components that describe major forms of variation for a given set of curves.

The FDA process required that the sampled f0 and intensity contours were represented by functions defined by a common time interval and number of measurement points (Ramsay & Silverman, 2005). As in Gubian et al. (2009), the sampled contours were interpolated using a 4th order B-spline basis function (cubic spline) with one knot per sample according to de Boor’s theorem (de Boor, 1972). The basis function defined the level of detail in which the set of functions represented the original contours.

Following the smoothing procedure, landmark registration was used to align peaks and valleys in the set of functions that have the same meaning, such as word and phoneme boundaries. The idea behind landmark registration is that functional data naturally vary not only vertically (e.g., changes in f0 or intensity), but also horizontally (e.g., phase variation), and in this project it was important to isolate f0 and intensity differences from phase variation by temporally aligning the functional data. In the database of expressive speech collected for this study, recordings of the same lexical content produced by different talkers (and even several repetitions by the same talker) had different overall durations, and the time onset of speaking was arbitrary. Landmark registration operates by identifying maxima, minima, and zero crossings of curves, and aligning those curves by warping the time domain (Ramsay & Silverman, 1997). Following the landmark registration procedure, mean f0 and intensity contours were computed separately for each speech type and emotion context.
The functional form of the mean contours was represented by a series of coefficients that statistically defined each sampled function. The coefficients of each of the individual contours were determined and entered into a principal components analysis to reveal detailed affective prosody production differences between the two groups, as well as individual differences within those groups. For example, consider the evoked phrase “I can’t believe this” used in this study. Two talkers may produce the phrase with a nearly equivalent f0 range, but the two f0 contours might be substantially different in terms of their shape and location of f0 and/or intensity peaks. One talker may produce the phrase in the happy emotion context with an increasing pattern of f0 on the word “believe” (e.g., I can’t believe this), while the other may produce the phrase with increased f0 on the word “can’t” (e.g., I can’t believe this). Both types of f0 changes were represented by coefficients describing the functional form of each f0 contour. To the extent that each type of f0 variation is common in the dataset, those patterns of f0 variation would be revealed by the functional PCA.

3.6.3 FDA results for f0

Functional PCA revealed patterns of f0 and intensity variation in the ASD and TD expressive speech data that were consistent with, yet more comprehensive than those reported in the acoustic analysis presented in section 3.5. Table 3.4 displays mean principal component scores and standard deviations for the first two principal components for each phrase and emotion context. The table includes functional ANOVA results and significance labels. The first two principal components explained an average of 73% (range across all analyses: 61%-85%) of the total variance in the functional data.
Table 3.4. Mean functional principal component scores and standard deviations for the analysis of f0 contours extracted from evoked phrases (neutral – p1 refers to phrase 1 produced in a neutral emotion context). Data are presented by talker group for each of the first two principal components. Total variance explained for each principal component (%var) and for the first two components combined (TotVarExp) are included. Functional ANOVA results and significance labels are included as well (* p < .05; ** p < .01; *** p < .001).

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th></th>
<th></th>
<th>PC2</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASD mean</td>
<td>sd</td>
<td>TD mean</td>
<td>sd</td>
<td>%var</td>
<td>ASD mean</td>
<td>sd</td>
<td>TD mean</td>
</tr>
<tr>
<td>neutral - p1</td>
<td>-0.05 0.72</td>
<td>0.05 0.70</td>
<td>59.4%</td>
<td>-0.04 0.50</td>
<td>0.04 0.42</td>
<td>25.2%</td>
<td>84.6%</td>
<td>0.17 0.49</td>
</tr>
<tr>
<td>neutral - p2</td>
<td>-0.17 0.60</td>
<td>0.17 0.55</td>
<td>45.2%</td>
<td>-0.07 0.54</td>
<td>0.07 0.42</td>
<td>29.9%</td>
<td>75.1%</td>
<td>0.40 0.00 ***</td>
</tr>
<tr>
<td>neutral - p3</td>
<td>0.05 0.72</td>
<td>-0.05 0.78</td>
<td>67.0%</td>
<td>0.02 0.42</td>
<td>-0.02 0.30</td>
<td>16.1%</td>
<td>83.1%</td>
<td>0.14 0.65</td>
</tr>
<tr>
<td>neutral - p4</td>
<td>-0.09 0.56</td>
<td>0.09 0.57</td>
<td>39.5%</td>
<td>-0.07 0.52</td>
<td>0.07 0.43</td>
<td>28.6%</td>
<td>68.1%</td>
<td>0.24 0.15</td>
</tr>
<tr>
<td>neutral - p5</td>
<td>-0.13 0.64</td>
<td>0.13 0.53</td>
<td>40.7%</td>
<td>-0.01 0.55</td>
<td>0.01 0.41</td>
<td>26.3%</td>
<td>67.0%</td>
<td>0.28 0.10</td>
</tr>
<tr>
<td>angry - p1</td>
<td>0.04 0.56</td>
<td>-0.04 0.58</td>
<td>49.4%</td>
<td>0.03 0.38</td>
<td>-0.03 0.37</td>
<td>21.2%</td>
<td>70.6%</td>
<td>0.10 0.82</td>
</tr>
<tr>
<td>angry - p2</td>
<td>-0.09 0.60</td>
<td>0.09 0.42</td>
<td>53.2%</td>
<td>0.01 0.41</td>
<td>-0.01 0.30</td>
<td>24.8%</td>
<td>78.0%</td>
<td>0.24 0.06</td>
</tr>
<tr>
<td>angry - p3</td>
<td>0.10 0.69</td>
<td>-0.10 0.68</td>
<td>57.3%</td>
<td>-0.09 0.43</td>
<td>0.09 0.36</td>
<td>20.0%</td>
<td>77.3%</td>
<td>0.31 0.04 *</td>
</tr>
<tr>
<td>angry - p4</td>
<td>0.10 0.57</td>
<td>-0.10 0.68</td>
<td>47.1%</td>
<td>-0.04 0.52</td>
<td>0.04 0.52</td>
<td>31.5%</td>
<td>78.6%</td>
<td>0.25 0.17</td>
</tr>
<tr>
<td>angry - p5</td>
<td>-0.03 0.43</td>
<td>0.03 0.48</td>
<td>40.8%</td>
<td>-0.02 0.44</td>
<td>0.02 0.24</td>
<td>25.0%</td>
<td>65.8%</td>
<td>0.17 0.25</td>
</tr>
<tr>
<td>happy - p1</td>
<td>0.08 0.54</td>
<td>-0.08 0.61</td>
<td>46.0%</td>
<td>0.02 0.46</td>
<td>-0.02 0.42</td>
<td>26.6%</td>
<td>72.6%</td>
<td>0.20 0.28</td>
</tr>
<tr>
<td>happy - p2</td>
<td>-0.16 0.51</td>
<td>0.16 0.30</td>
<td>39.9%</td>
<td>-0.02 0.40</td>
<td>0.02 0.37</td>
<td>30.0%</td>
<td>69.9%</td>
<td>0.34 0.00 ***</td>
</tr>
<tr>
<td>happy - p3</td>
<td>0.14 0.59</td>
<td>-0.14 0.61</td>
<td>49.5%</td>
<td>-0.04 0.43</td>
<td>0.04 0.43</td>
<td>24.6%</td>
<td>74.1%</td>
<td>0.35 0.01 ***</td>
</tr>
<tr>
<td>happy - p4</td>
<td>0.05 0.57</td>
<td>-0.05 0.53</td>
<td>34.3%</td>
<td>-0.11 0.44</td>
<td>0.11 0.59</td>
<td>31.8%</td>
<td>66.1%</td>
<td>0.28 0.05</td>
</tr>
<tr>
<td>happy - p5</td>
<td>-0.07 0.54</td>
<td>0.07 0.48</td>
<td>36.0%</td>
<td>0.18 0.46</td>
<td>-0.18 0.35</td>
<td>27.4%</td>
<td>63.4%</td>
<td>0.45 0.00 ***</td>
</tr>
<tr>
<td>interested - p1</td>
<td>0.11 0.53</td>
<td>-0.10 0.65</td>
<td>46.0%</td>
<td>0.04 0.55</td>
<td>-0.04 0.36</td>
<td>27.4%</td>
<td>73.4%</td>
<td>0.31 0.05</td>
</tr>
<tr>
<td>interested - p2</td>
<td>-0.27 0.71</td>
<td>0.27 0.32</td>
<td>51.5%</td>
<td>0.08 0.50</td>
<td>-0.08 0.36</td>
<td>27.5%</td>
<td>79.0%</td>
<td>0.57 0.00 ***</td>
</tr>
<tr>
<td>interested - p3</td>
<td>-0.06 0.66</td>
<td>0.06 0.70</td>
<td>48.9%</td>
<td>-0.01 0.48</td>
<td>0.01 0.46</td>
<td>22.8%</td>
<td>71.7%</td>
<td>0.22 0.34</td>
</tr>
<tr>
<td>interested - p4</td>
<td>-0.24 0.71</td>
<td>0.24 0.49</td>
<td>44.5%</td>
<td>-0.15 0.47</td>
<td>0.15 0.52</td>
<td>28.2%</td>
<td>72.7%</td>
<td>0.57 0.00 ***</td>
</tr>
<tr>
<td>interested - p5</td>
<td>0.07 0.51</td>
<td>-0.07 0.52</td>
<td>32.3%</td>
<td>-0.15 0.51</td>
<td>0.15 0.42</td>
<td>29.1%</td>
<td>61.4%</td>
<td>0.46 0.00 ***</td>
</tr>
<tr>
<td>sad - p1</td>
<td>0.02 0.63</td>
<td>-0.02 0.63</td>
<td>51.5%</td>
<td>-0.08 0.42</td>
<td>0.08 0.49</td>
<td>27.7%</td>
<td>79.2%</td>
<td>0.17 0.47</td>
</tr>
<tr>
<td>sad - p2</td>
<td>-0.02 0.50</td>
<td>0.02 0.35</td>
<td>36.9%</td>
<td>-0.13 0.43</td>
<td>0.13 0.33</td>
<td>32.9%</td>
<td>69.8%</td>
<td>0.27 0.00 **</td>
</tr>
<tr>
<td>sad - p3</td>
<td>0.02 0.69</td>
<td>-0.02 0.74</td>
<td>59.6%</td>
<td>-0.02 0.45</td>
<td>0.02 0.35</td>
<td>19.0%</td>
<td>78.6%</td>
<td>0.07 0.96</td>
</tr>
<tr>
<td>sad - p4</td>
<td>-0.18 0.66</td>
<td>0.18 0.56</td>
<td>46.7%</td>
<td>-0.02 0.54</td>
<td>0.02 0.46</td>
<td>29.2%</td>
<td>75.9%</td>
<td>0.37 0.01 **</td>
</tr>
<tr>
<td>sad - p5</td>
<td>0.13 0.48</td>
<td>-0.13 0.63</td>
<td>45.1%</td>
<td>-0.08 0.43</td>
<td>0.08 0.37</td>
<td>23.0%</td>
<td>68.1%</td>
<td>0.34 0.01 ***</td>
</tr>
</tbody>
</table>

Figure 3.12 shows an example plot of the most common forms of f0 variation for the phrase “why did you do that?” produced in the happy emotion context. Repetitions were collapsed given that no main effect for f0 range by repetition number was found (see section...
3.5). The y-axis labeled “Harmonic 1” corresponds to the eigenvalue scores for the first principal component, corresponding to the most important mode of variation in the functional data for this phrase and emotion context.

Figure 3.12. Functional principal component analysis results (principal component 1) for productions of the evoked phrase “why did you do that” produced in the happy emotion context. The solid line represents the functional mean for both groups. The dashed line (-) represents the shape of the mean f0 contour when the first principal component (PC1) scores were negative, (+) describes the shape of the f0 contour when the PC1 scores were positive.

For each analysis, similar principal component plots were available to compare individual contours and talker group mean contours to the overall mean contour. In this example, the first principal component explains 39.9% of the variance in the functional data, and shows that the
most common form of variation in the data corresponds to an early or late peak in the f0 contours.

Figure 3.13. Example f0 contour FDA plot for the evoked phrase “why did you do that” produced in a happy emotion context. Individual contours are shown as faint dotted lines and mean contours are shown as a solid black line (the mean contour for both groups), solid green line (TD talkers) and a solid red line (talkers with ASD).

Table 3.4 shows that the mean PC1 score for phrase 2 produced in a happy context was negative (-0.161) for talkers with ASD and positive for TD talkers (0.161). In the example provided in Figure 3.12, the negative PC1 score obtained for talkers with ASD corresponded to an f0 contour with an early peak followed by a drop in f0 through the rest of the phrase. In contrast, the positive PC1 score reflected by TD talkers began at a lower level and peaked later in
the target phrase (“why did you do that”). Further investigation revealed that TD talkers had maximum f0 values for the fourth word in the phrase (“do”), and produced the phrase with more consistent f0 contour shapes compared to talkers with ASD.

Figure 3.13 displays group and overall mean functions for happy productions of phrase 2, using the same data presented in the first principal component shown in Figure 3.12. The overall mean is shown as a solid black line, and the mean functions for ASD talkers (red line) and TD talkers (green line) are shown separately. Faint dotted lines representing the individual contours are displayed in their respective group colors. Figure 3.13 shows that talkers with ASD produced the phrase with greater f0 variability compared to TD talkers, who yielded a visibly more consistent pattern of f0 production.

Review of Figures 3.12 and 3.13 is instructive in explaining the pattern of group differences identified in the functional PCA, where the first principal component corresponded to an early peak/late peak distinction between the two groups. Talkers with ASD were more likely to produce the phrase with an early f0 peak, and the individual and mean contours in Figures 3.12 and 3.13 show the early peak pattern. The solid red line in Figure 3.13 shows the mean ASD function peaking early in the phrase, while the solid green line showing the TD talker group peaks late and matches the late peak pattern for TD talkers in Figure 3.12. In this example, the consistency of the TD talker f0 contours is clearly shown – most (but not all) of the individual contours follow the late peak pattern. In comparison, greater variability was found in f0 contours produced by talkers with ASD.
3.6.4 FDA results for intensity

The FDA performed on intensity contours extracted from evoked phrases highlighted types of intensity variation between the two talker groups. The FDA results complement the analysis of mean intensity presented in Section 3.5. Table 3.5 displays mean principal component scores and standard deviations for the first two principal components for each phrase and emotion context. The table includes functional ANOVA results and significance labels. The first two principal components explained an average of 79% of the total variance in the functional data (range across all analyses: 71%-87%).

The data in Table 3.5 indicate that the first principal component explained approximately half of the variability in the functional data for each analysis, and the second component explained approximately 30% of the variance. Figure 3.14 shows an example of the most common form of variation for angry productions of the phrase “I can’t believe this.” The y-axis labeled “Harmonic 1” corresponds to the eigenvalue scores for the first principal component, corresponding to the most important mode of variation in the functional data for this phrase and emotion context. In this example, the first principal component explains 47.5% of the variance in the functional data, and shows that the most common form of variation in the data corresponds to an early or late peak in the intensity contours. A review of the corresponding analysis in Table 3.5 (angry – p3) shows that the ANOVA comparing the difference between the mean function for TD talkers and talkers with ASD was significant, and that talkers with ASD had a negative value for PC1. The intensity contours produced by talkers with ASD were more likely to resemble the dashed (-) lines in Figure 3.14, corresponding to a late peak in intensity.
Table 3.5. Mean functional principal component scores and standard deviations for the analysis of intensity contours extracted from evoked phrases (i.e., neutral – p1 refers to phrase 1 produced in a neutral emotion context). Data are presented by talker group for each of the first two principal components. Total variance explained for each principal component (%var) and for the first two components combined (TotVarExp) are included. Functional ANOVA results and significance labels are included as well (* p < .05; ** p < .01; *** p < .001).

<table>
<thead>
<tr>
<th>Group</th>
<th>PC1 mean</th>
<th>PC1 sd</th>
<th>TD mean</th>
<th>TD sd</th>
<th>%var</th>
<th>PC2 mean</th>
<th>PC2 sd</th>
<th>TD mean</th>
<th>TD sd</th>
<th>%var</th>
<th>Tot Var Exp</th>
<th>Group Difference</th>
<th>ANOVA</th>
<th>stat.</th>
<th>p</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutral - p1</td>
<td>0.12</td>
<td>0.44</td>
<td>-0.12</td>
<td>0.42</td>
<td>43.1%</td>
<td>-0.03</td>
<td>0.46</td>
<td>0.03</td>
<td>0.32</td>
<td>33.5%</td>
<td>76.6%</td>
<td>0.28</td>
<td>0.01</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>neutral - p2</td>
<td>0.02</td>
<td>0.44</td>
<td>-0.02</td>
<td>0.40</td>
<td>59.8%</td>
<td>0.09</td>
<td>0.33</td>
<td>-0.09</td>
<td>0.19</td>
<td>27.1%</td>
<td>86.9%</td>
<td>0.19</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neutral - p3</td>
<td>-0.06</td>
<td>0.35</td>
<td>0.06</td>
<td>0.36</td>
<td>52.4%</td>
<td>0.09</td>
<td>0.28</td>
<td>-0.09</td>
<td>0.17</td>
<td>25.0%</td>
<td>77.5%</td>
<td>0.23</td>
<td>0.00</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>neutral - p4</td>
<td>-0.01</td>
<td>0.47</td>
<td>0.01</td>
<td>0.43</td>
<td>45.7%</td>
<td>0.05</td>
<td>0.40</td>
<td>-0.05</td>
<td>0.38</td>
<td>35.0%</td>
<td>80.7%</td>
<td>0.13</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neutral - p5</td>
<td>-0.07</td>
<td>0.47</td>
<td>0.07</td>
<td>0.47</td>
<td>53.6%</td>
<td>0.04</td>
<td>0.35</td>
<td>-0.04</td>
<td>0.33</td>
<td>27.6%</td>
<td>81.2%</td>
<td>0.16</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>angry - p1</td>
<td>0.11</td>
<td>0.51</td>
<td>-0.11</td>
<td>0.46</td>
<td>47.1%</td>
<td>-0.09</td>
<td>0.50</td>
<td>0.09</td>
<td>0.34</td>
<td>36.0%</td>
<td>83.1%</td>
<td>0.30</td>
<td>0.01</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>angry - p2</td>
<td>-0.02</td>
<td>0.39</td>
<td>0.02</td>
<td>0.38</td>
<td>46.0%</td>
<td>0.12</td>
<td>0.41</td>
<td>-0.12</td>
<td>0.14</td>
<td>34.4%</td>
<td>80.4%</td>
<td>0.26</td>
<td>0.01</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>angry - p3</td>
<td>-0.10</td>
<td>0.39</td>
<td>0.10</td>
<td>0.30</td>
<td>47.5%</td>
<td>0.08</td>
<td>0.33</td>
<td>-0.08</td>
<td>0.20</td>
<td>30.1%</td>
<td>77.6%</td>
<td>0.26</td>
<td>0.00</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>angry - p4</td>
<td>0.10</td>
<td>0.46</td>
<td>-0.10</td>
<td>0.40</td>
<td>49.7%</td>
<td>-0.04</td>
<td>0.40</td>
<td>0.04</td>
<td>0.29</td>
<td>31.0%</td>
<td>80.7%</td>
<td>0.24</td>
<td>0.04</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>angry - p5</td>
<td>-0.03</td>
<td>0.45</td>
<td>0.03</td>
<td>0.46</td>
<td>53.4%</td>
<td>0.00</td>
<td>0.34</td>
<td>0.00</td>
<td>0.31</td>
<td>27.5%</td>
<td>81.0%</td>
<td>0.10</td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>happy - p1</td>
<td>0.09</td>
<td>0.50</td>
<td>-0.09</td>
<td>0.41</td>
<td>50.3%</td>
<td>-0.08</td>
<td>0.40</td>
<td>0.08</td>
<td>0.21</td>
<td>24.9%</td>
<td>75.2%</td>
<td>0.25</td>
<td>0.03</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>happy - p2</td>
<td>-0.03</td>
<td>0.42</td>
<td>0.03</td>
<td>0.44</td>
<td>52.8%</td>
<td>0.10</td>
<td>0.39</td>
<td>-0.10</td>
<td>0.16</td>
<td>28.1%</td>
<td>80.8%</td>
<td>0.24</td>
<td>0.03</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>happy - p3</td>
<td>-0.07</td>
<td>0.34</td>
<td>0.07</td>
<td>0.39</td>
<td>52.6%</td>
<td>0.09</td>
<td>0.29</td>
<td>-0.09</td>
<td>0.18</td>
<td>25.6%</td>
<td>78.2%</td>
<td>0.24</td>
<td>0.01</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>happy - p4</td>
<td>0.00</td>
<td>0.41</td>
<td>0.00</td>
<td>0.38</td>
<td>41.7%</td>
<td>0.09</td>
<td>0.41</td>
<td>-0.09</td>
<td>0.30</td>
<td>36.4%</td>
<td>78.1%</td>
<td>0.19</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>happy - p5</td>
<td>0.01</td>
<td>0.50</td>
<td>-0.01</td>
<td>0.47</td>
<td>46.9%</td>
<td>-0.01</td>
<td>0.39</td>
<td>0.01</td>
<td>0.30</td>
<td>24.3%</td>
<td>71.2%</td>
<td>0.13</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interested - p1</td>
<td>0.06</td>
<td>0.52</td>
<td>-0.06</td>
<td>0.35</td>
<td>45.4%</td>
<td>-0.03</td>
<td>0.45</td>
<td>0.03</td>
<td>0.28</td>
<td>32.6%</td>
<td>78.0%</td>
<td>0.26</td>
<td>0.02</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>interested - p2</td>
<td>0.02</td>
<td>0.34</td>
<td>-0.02</td>
<td>0.47</td>
<td>52.7%</td>
<td>0.09</td>
<td>0.40</td>
<td>-0.09</td>
<td>0.17</td>
<td>31.9%</td>
<td>84.6%</td>
<td>0.20</td>
<td>0.04</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>interested - p3</td>
<td>-0.10</td>
<td>0.33</td>
<td>0.10</td>
<td>0.37</td>
<td>40.0%</td>
<td>0.13</td>
<td>0.39</td>
<td>-0.13</td>
<td>0.20</td>
<td>33.9%</td>
<td>73.9%</td>
<td>0.32</td>
<td>0.00</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>interested - p4</td>
<td>0.07</td>
<td>0.44</td>
<td>-0.07</td>
<td>0.48</td>
<td>48.7%</td>
<td>0.14</td>
<td>0.35</td>
<td>-0.14</td>
<td>0.31</td>
<td>29.5%</td>
<td>78.1%</td>
<td>0.33</td>
<td>0.00</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>interested - p5</td>
<td>-0.05</td>
<td>0.47</td>
<td>0.05</td>
<td>0.47</td>
<td>44.6%</td>
<td>0.05</td>
<td>0.42</td>
<td>-0.05</td>
<td>0.33</td>
<td>28.8%</td>
<td>73.3%</td>
<td>0.18</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sad - p1</td>
<td>0.05</td>
<td>0.47</td>
<td>-0.05</td>
<td>0.44</td>
<td>48.9%</td>
<td>-0.02</td>
<td>0.42</td>
<td>0.02</td>
<td>0.32</td>
<td>32.0%</td>
<td>80.8%</td>
<td>0.14</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sad - p2</td>
<td>-0.07</td>
<td>0.42</td>
<td>0.07</td>
<td>0.40</td>
<td>52.1%</td>
<td>-0.10</td>
<td>0.38</td>
<td>0.10</td>
<td>0.18</td>
<td>29.3%</td>
<td>81.4%</td>
<td>0.26</td>
<td>0.00</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sad - p3</td>
<td>-0.14</td>
<td>0.41</td>
<td>0.14</td>
<td>0.24</td>
<td>50.1%</td>
<td>0.08</td>
<td>0.36</td>
<td>-0.08</td>
<td>0.22</td>
<td>36.0%</td>
<td>86.1%</td>
<td>0.32</td>
<td>0.00</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sad - p4</td>
<td>-0.01</td>
<td>0.48</td>
<td>0.01</td>
<td>0.48</td>
<td>47.9%</td>
<td>0.10</td>
<td>0.35</td>
<td>-0.10</td>
<td>0.39</td>
<td>30.2%</td>
<td>78.0%</td>
<td>0.20</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sad - p5</td>
<td>-0.05</td>
<td>0.58</td>
<td>0.05</td>
<td>0.53</td>
<td>60.5%</td>
<td>0.01</td>
<td>0.36</td>
<td>-0.01</td>
<td>0.30</td>
<td>21.5%</td>
<td>82.0%</td>
<td>0.15</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.15 displays group and overall mean functions for angry productions of the phrase “I can’t believe this,” which corresponds to the same data presented in Figure 3.14. The overall mean is shown as a solid black line, and the mean intensity functions for ASD talkers
(red line) and TD talkers (green line) are shown separately. Faint dotted lines representing the individual contours are displayed in their respective group colors. Figure 3.15 reveals that talkers with ASD produced the phrase with greater variability in intensity compared to TD talkers, who exhibited a visibly more consistent pattern.

Figure 3.14. Functional principal component analysis results (principal component 1) for productions of the evoked phrase “I can’t believe this” produced in the angry emotion context. The solid line represents the functional mean for both groups. The dashed line (-) represents the shape of the mean intensity contour when the first principal component (PC1) scores were negative. (+) describes the shape of the intensity contour when the PC1 scores were positive.
Review of Figures 3.14 and 3.15 is instructive in explaining the pattern of group differences in intensity identified in the functional PCA, where the first principal component corresponded to an early peak/late peak distinction between the two groups. Talkers with ASD were more likely to produce the phrase with a late peak in intensity, as shown by the solid red line in Figure 3.15. In contrast, the solid green line showing the mean contour for the TD group peaks early, followed by a gradual drop in intensity. In this example, the pattern of more consistent intensity production from TD talkers is clear – most (but not all) of the individual
contours follow the early peak pattern, while a more inconsistent pattern is apparent in the individual contours produced by talkers with ASD.

3.7 Discussion

This chapter reported a series of acoustic analyses performed to compare affective prosody production in talkers with ASD and TD controls. Expressive speech recordings produced in five emotion contexts consisting of evoked phrases, VCV syllables and scripted dialogue sessions were analyzed to investigate potential talker group differences in acoustic properties relevant for communication of affective prosody. ANOVAs were performed on summary measures of f0 range, mean intensity and duration, resulting in the general finding that talkers with ASD produced the expressive speech recordings with increased variability in acoustic properties relevant for communication of affective prosody. To further investigate modes of variation within the utterances, additional tests were performed on the f0 and intensity contours extracted from evoked phrases using functional data analysis. Prior research has shown that talkers with ASD produce prosody with increased variability in f0 compared to control groups (Fosnot & Jun, 1999; Nadig & Shaw, 2012). The current acoustic analysis is consistent with those studies, and provides additional evidence of inconsistent patterns of affective prosody production in a sample of adult males with ASD.

A key finding was that talkers with ASD produced affective prosody with increased f0 variability, compared to controls, in each emotion context except neutral for all speech types with the exception of happy VCV syllables. This suggests that increased f0 variability is an emotion-specific property of affective prosody production in talkers with ASD rather than a general characteristic of all speech, and is only associated with emotion contexts which contain
modulation of the f0 contour. Additionally, the greatest differences in f0 range between talkers with ASD and controls were found in the high-arousal emotion contexts (e.g., angry and happy), particularly for the evoked phrases and scripted dialogues (e.g., longer speech segments).

Production of f0 is known to be closely linked to overall arousal level, and the current findings suggest that talkers with ASD engaged in the evoked phrases and prompted VCV tasks with greater overall arousal relative to TD talkers.

The analysis of intensity revealed that speech produced by talkers with ASD contained higher mean intensity levels and intensity variability in all five emotion contexts. In the evoked phrases, both talker groups increased intensity from the first to third repetitions in the high arousal emotions (angry and happy), and decreased intensity for the sad, interested and neutral (low arousal) contexts. Figure 3.3 displays that result, and shows that talkers with ASD engaged in the task with greater overall intensity in each emotion context. This pattern of production was consistent for the evoked phrases and VCV syllables, but the result did not hold up in the scripted dialogue sessions, where no difference was found in intensity between the two talker groups.

The analysis of duration revealed that talkers with ASD produced VCV syllables and evoked phrases with longer durations compared to those produced by TD talkers. The results show expected patterns of acoustic modulations consistent with those previously reported in the literature (Scherer, 1979, 1986; Murray and Arnott, 1993; for a review see Juslin and Laukka, 2003), in that the sad and interested contexts are low-arousal emotions and were characterized by longer durations, whereas the angry and happy contexts are high-arousal emotions and were
accompanied by shorter durations. As in the analysis of intensity, however, the scripted dialogue sessions contained no differences in duration between the talker groups.

The findings support the prediction that acoustic properties relevant to affective prosody produced by talkers with ASD would display more variable patterns of production compared to those produced by TD talkers. In contrast with prior research in this area using semi-spontaneous or spontaneous speech, the current analyses were performed on recordings in which each talker produced the same verbal content, which allowed a more comprehensive analysis of modes of variation in affective prosody between the two groups. In addition, increased f0 and intensity variability were found in lexically-ambiguous phrases and VCV syllables. This suggests that inconsistent patterns of affective prosody production in talkers with ASD are not dependent upon the verbal content in the speech produced but rather may part of a more general deficit in producing emotion in a wide range of speech types.

Overall, the acoustic analysis suggests that talkers with ASD and controls exhibit similar emotion-specific patterns of prosody production, but that talkers with ASD produce what could be considered as more extreme versions of the those properties, even though the lexical content of the speech task was neutral in the evoked phrases and VCV syllables. A potential explanation for the affective prosody production differences reported in this chapter involves diminished social motivation and attention in ASD. Research has suggested a link between production and perception of affective prosody, and inattention to social norms for emotional expression is a potential cause for motivational differences in prosody production, due to the inability to utilize those norms during speech production. Nadig and Shaw (2012) also found greater f0 variability
in productions of prosody by talkers with ASD, and concluded that deficits in auditory processing may impact one’s own speech production, leading to abnormal utterances.

Diminished social motivation is part of a larger theoretical framework showing auditory processing deficits in individuals with ASD (O'Connor, 2012). A considerable amount of behavioral research has reported atypical patterns of social orientation and social attention (Klin, 1991; Kuhl et al., 2005; Dawson et al., 2004), particularly to spoken language, that occur as early as 8-10 months of age (Werner et al., 2000). Meanwhile, divergent patterns of affective prosody production are suggested to impact listener perceptions of oddness (Van Bourgondien & Woods, 1992; Paul et al., 2005), and gaining a better understanding of these production differences is important to enhancing social communication skills for talkers with ASD.

A potential reason for the absence of group effects in intensity and duration in the scripted dialogue sessions is that talkers with ASD may have been aided by the lexical content contained in the phrases. In those recordings, the verbal content matched the target emotion context (i.e., “I just won the lottery” for the happy context), and adult talkers may have been able to utilize prior learning about expected patterns of affective prosody production based on that content alone. That explanation, however, does not account for the result that, like the evoked phrases and VCV syllables, f0 range was still higher in the dialogue phrases. While it is clear that talkers with ASD produce affective prosody with less consistency compared to controls, further research is warranted to better understand links between the lexical content and type of speech tasks and changes in acoustic properties associated with affective prosody. Future investigations involving different stimulus types and elicitation methods will provide insight into
acoustic variability and the role of lexical ambiguity in productions of affective prosody by talkers with ASD.
CHAPTER 4
PERCEPTION OF AFFECTIVE PROSODY

4.1 Introduction

What is the perceptual impact of increased variability in affective prosody produced by talkers with ASD? This chapter describes a series of listening experiments conducted to examine the role of specific voice properties involved in perception of affective prosody, and to compare perception of affective prosody produced by individuals with autism (ASD) and typically-developing (TD) controls. In each experiment, evoked expressive phrases taken from the expressive speech database were presented to listeners, who were asked to identify the emotion context and rate the level of naturalness in the emotional content in each stimulus. In the acoustic analysis presented in Section 3.5, reliable talker group differences in f0 and intensity were found in the expressive phrases, consistent with previously reported results (Nadig & Shaw, 2012; Hubbard & Trauner, 2007).

Based on the hypothesis that divergent patterns of affective prosody production would lead to inconsistent patterns of listener perception, the first two experiments were designed to investigate the perceptual relevance of increased f0 and intensity variability found in the recordings from talkers with ASD. Inconsistent modes of f0 and intensity production in talkers with ASD may contribute to perceptions of social oddness from communication partners and negatively impact social communication outcomes in individuals with ASD. In Experiment 1, f0-modified and f0-unmodified stimuli were presented to TD listeners to examine the perceptual
consequence of inconsistent patterns of f0 produced by talkers with ASD. In the modified condition, the f0 contour of each test stimulus was replaced by the mean contour for the TD group, resulting in reduced f0 variability. In Experiment 2, intensity-modified and intensity-unmodified stimuli were presented to TD listeners to assess the perceptual impact of inconsistent patterns of intensity production in affective prosody. In the modified condition, the intensity contour of each test stimulus was replaced by the mean contour for the TD group, resulting in reduced mean intensity and reduced intensity variability. In Experiment 3, unmodified versions of the test stimuli were presented to listeners with ASD and TD controls to assess potential differences in patterns of listener responses between the two groups, and potential interactions between the talker and listener groups.

4.1.1 Method and procedure common to all experiments

In each experiment, a subset of evoked phrases taken from the expressive speech database was presented to listeners in emotion identification tasks. Of the speech types recorded for the database, evoked phrases were the most suitable for use in listening tests because they contained real words (as opposed to the VCV syllables), and given the ambiguous lexical content of the phrases (as opposed to the scripted dialogue sessions), perception of emotional content in the phrases was dependent upon the ability to access prosodic cues in the test stimuli. The listening tests included 330 trials (360 for Experiment 3); in each trial participants listened to the test phrase over headphones, judged the emotion context by clicking an emotion button, then rated the level of naturalness of the emotional content in the phrase. Emotion labels were collected using a five-item forced-alternative test, and naturalness ratings were collected on a slider-type scale ranging from 0 to 10.
To ensure listening test stimuli were representative of the five emotion contexts recorded, and to limit testing to approximately one hour, a subset of evoked phrases was selected based on previously reported patterns of voice property modulations associated with each emotion context. (e.g., Scherer, 1979, 1986; Murray and Arnott, 1993; for a review see Juslin and Laukka, 2003). 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, based on reports of increased f0 variability in those contexts (Scherer, 1986, 1991; Murray & Arnott, 1993, for a review see Juslin & Laukka, 2003) and consistent with the analysis presented in Chapter 3. For the angry context, phrases with the highest median intensity were identified based on the result that angry speech contained the highest levels of intensity in the database recordings and that increased intensity (sometimes described as loudness) is a key marker for angry speech reported in the literature (Scherer, 1986; Murray & Arnott, 1993). Sad phrases with the greatest negative f0 slope from the beginning to the end of the phrase were selected, based on previously reported results showing a downward f0 trajectory in sad speech (Juslin & Laukka, 2003). Neutral recordings with the smallest f0 range were selected, corresponding to a relatively flat f0 contour typical of neutral speech. Additional constraints were placed on the selection of stimuli to ensure the number of stimuli produced by talkers with ASD and TD and in each emotion context was balanced. All ambiguous phrase types and talkers were included in the set of test stimuli.

During testing, an audio presentation of the test phrase was followed by the display of five emotion buttons on a computer screen to capture the emotion context judgment for each stimulus. The button labels included the five recorded contexts discussed in Chapter 2: “neutral”; “angry”; “happy”; “interested”; and “sad”, and the buttons appeared in random order per listener
to eliminate the possibility that the position of a button on the screen would impact the overall pattern of responses. Prior to beginning the identification test, listeners were instructed to “listen carefully to each stimulus and identify which emotion context you heard by clicking the appropriate button on the computer screen” and that “some of the emotions are difficult to distinguish, but please click on the button that most closely matches the emotion you heard.” Following each emotion context decision, a naturalness rating was obtained using a slider-type scale ranging from 0 (least natural sounding) to 10 (most natural sounding). Listeners were read the following instructions: “after you select an emotion context the system will prompt you to rate how natural the phrase you heard was. For the purposes of this experiment, naturalness is defined as the degree to which the emotional aspects of the phrase would be encountered in everyday expressive communication. The naturalness rating is on a slider-type scale from 0 to 10, and your task will be to use the computer mouse to slide the marker to indicate a low naturalness rating on the left end of the slider or a high naturalness rating on the right side, or somewhere in between.”

Each testing session began with obtaining informed consent, a brief questionnaire, a brief hearing test, 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). Prior to each session a brief hearing test was performed to confirm that listeners were able to hear and respond to a set of pure tones presented at different frequencies at levels corresponding to standard audiometric thresholds for listeners with normal hearing. Each ear was tested separately, and participants were asked to respond
when they heard a tone by raising a hand. All participants were able to pass the hearing test without trouble. The practice session was designed to familiarize listeners with the phrases used and with the user interface; it included stimuli not in the main experiment. Stimuli were presented diotically through Sennheiser HD-598 headphones using Tucker-Davis System 3 and RP2.1 hardware. All stimulus conditions were randomized and presented using custom Matlab scripts. 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.

4.2 Experiment 1: The role of f0 in affective prosody perception

4.2.1 Introduction

In the acoustic analysis of affective prosody presented in Section 3.5.1, talkers with ASD produced evoked phrases in each of the non-neutral emotion contexts with increased f0 range compared to TD talkers. The overall pattern of results revealed greater f0 range and f0 variability (f0 SD) in the recordings produced by ASD talkers, whereas TD talkers produced the phrases using more consistent patterns of f0 production. This pattern was evident in the ANOVA on f0 range by talker group, and in the functional data analysis results which revealed more detailed modes of variation in the mean f0 contours for each talker group. Thus, Experiment 1 was designed to study the impact of talker group-based differences in f0 production on listener perception of affective prosody. Original and f0-modified versions of the phrases were presented to naïve listeners and emotion category responses and naturalness ratings were collected.
F0 corresponds to the rate at which the vocal folds open and close and is measured by the number of open-close cycles per second, or Hertz (Hz). It is strongly correlated with perceived voice pitch, and as such it is a basic acoustic property involved in human speech. F0 varies dynamically in running speech to signal not only expressive content but other forms of prosody and indexical information as well. It has been established that F0 plays an important role in communication of emotion in speech, and systematic modulations to F0 that accompany different emotion contexts have been documented by a growing number of studies (e.g., Scherer, 1979, 1986; Murray and Arnott, 1993; for a review see Juslin and Laukka, 2003). F0 has been called the best indicator of the emotional content in an utterance. In Williams and Stevens (1972), actors expressed four different emotion contexts (neutral, anger, fear, and sorrow) and comparisons were made between the emotions using spectrographic analysis. The authors reported that deviation in F0, compared to the smooth and slowly varying patterns of F0 production typical of neutral speech, offered the best indicator of a talker’s intended emotion.

To investigate the perceptual impact of differences in F0 production found in the ASD talker group, phrases were presented to listeners in two conditions: unmodified versions of the original recordings and F0-modified versions of the same group of stimuli. In the F0-modified stimuli, the F0 contour was replaced by the average contour for the TD talker group. The design provided the ability to compare perception between the two talker groups in the unmodified stimuli and to determine if reducing the F0 variability in the stimuli would cause the two groups to be perceived more similarly.
4.2.2 Method and procedure

The listening test to examine affective prosody perception for talkers with ASD and TD controls, and the impact of f0-related production differences included 360 trials; in each trial listeners judged the emotion context and level of perceived naturalness in the emotional content of each phrase. The test stimuli included 180 recordings per condition (f0-modified and f0-unmodified), and the stimuli were divided evenly by emotion context, resulting in 36 stimuli per emotion context by condition group. The f0-modified and f0-unmodified stimuli were derived of the same set of recordings, with the experimental manipulation applied to the f0-modified group.

The process to manipulate f0 began by extracting the f0 contour from each recording using STRAIGHT, a high-quality speech analysis and resynthesis tool that generates synthetic speech that is nearly indistinguishable from the original recordings (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999). STRAIGHT analyzes input speech into separate components related to the voicing source (F0 and aperiodicity) and the vocal tract transfer function (spectrum envelope). Following the f0 estimation, the contours taken from the TD group were processed using functional data analysis (Ramsay & Silverman, 2005) to obtain the mean f0 contour for the TD group. Mean contours were calculated separately for each phrase and emotion context, given that patterns of f0 production are expected to vary when speaking in different emotional contexts and/or using different lexical content.

FDA requires that the sampled f0 contours were defined by a common time interval and number of measurement points (Ramsay & Silverman, 2005). Therefore, each contour was time-normalized and represented by 200 slices of corresponding to individual f0 measurements. As in Gubian et al. (2009), the sampled contours were interpolated using a fourth order B-spline basis
function (cubic spline) with one knot per sample according to de Boor’s theorem (de Boor, 1972). The basis function defined the level of detail in which the set of functions represented the original contours.

Following the smoothing procedure, landmark registration was used to align peaks and valleys in the set of functions that have the same meaning, such as word and phoneme boundaries. In the set of evoked phrases used for this experiment, recordings of the same lexical content produced by different talkers contained different overall durations, and landmark registration was used to align important features of each contour prior to computing the mean function to be used in the synthesis step. Following the landmark registration procedure, the mean f0 contour was computed separately for each phrase and emotion context.

To create f0-modified test stimuli, the f0 contour of each phrase produced by both groups of talkers was replaced by the mean f0 contour for the TD group. During analysis, STRAIGHT separates the input speech into the f0 contour and the spectrum envelope, which enabled synthesis of stimuli in which only the f0 contour was manipulated while leaving the spectrum envelope and other properties unchanged. To ensure that any perceptual patterns found were not the result of the synthesis procedure, f0-unmodified stimuli were resynthesized in the same manner (but without the f0 modification), creating 180 f0-modified and 180 f0-unmodified versions of the same recordings. Figure 4.1 shows f0 contours for the phrase “what, what do you mean?” produced in the happy context. The figure shows the effect of f0-modification for a portion of the happy phrases that were presented to listeners. For additional details about the listening test method and procedure, please see Section 4.1.
Figure 4.1. Example f0 contours for the evoked phrase “what, what do you mean?” produced in the happy context. Individual contours are shown for each talker group and condition for phrases presented to listeners in the emotion identification test.
4.2.3 Participants

Twenty typically-developing listeners served as participants in the emotion identification test (19 females, 1 male, mean age 22.2 years). All listeners were native speakers of American English with no reported speech or hearing problems, were recruited from the UTD Behavioral and Brain Sciences participant pool and received research participation credit as compensation. Each listener self-reported no history of being diagnosed with autism, and completed the Broad Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007) to assess their tendency to display traits consistent the Broad Autism Phenotype (BAP). All participants had an overall BAPQ score below the increased specificity cutoff (3.55) suggested by Sasson and colleagues (2013), which suggests they did not display traits consistent with BAP. For additional details about the method and procedure, see Section 4.1.1

4.2.4 Data analysis

Each listener provided emotion category judgments and naturalness ratings for 360 phrases (72 per emotion context, divided evenly by experimental condition). Emotion recognition accuracy (ERA) was calculated by dividing the number of correctly identified phrases by the number produced in that context. Mean naturalness ratings were calculated by averaging individual scores across items. The data were summarized to enable comparisons by experimental condition, talker group and emotion context. ERA and naturalness were treated as dependent variables in separate mixed effects regression analyses, and talker group, emotion context, phrase and repetition were used as independent variables. Listeners were treated as a repeated random variable given that each listener provided multiple emotion context judgments
and naturalness ratings. The R statistical computing software (R Core Team, 2015) and the “lme4” package (Bates et al., 2015) were used to perform the ERA and naturalness analyses.

### 4.2.5 Hypotheses

Research has suggested that abnormal speech production can contribute to impressions of social oddness when communicating with someone with ASD (Van Bourgondien and Woods, 1992; Paul et al., 2005). An acoustic basis for impressions of social oddness was presented in Chapter 3; talkers with ASD produced the set of evoked phrases with greater variability in f0 and with higher intensity compared to the TD talker group. Based on those results and previously reported findings suggesting that f0 is the most important voice cue for perception of affective prosody, the predicted result was that listener emotion recognition accuracy (ERA) would decrease for phrases produced by talkers with ASD compared to those produced by TD talkers when responding to the unmodified stimuli. In addition, to the extent that increased variation in f0 leads to lower ERA scores, it is predicted that ERA for the f0-modified stimuli will not differ for the two talker groups.

### 4.2.6 Results

For ERA, main effects were found for condition \( (F (1, 7161) = 32.925, p < .001) \) and emotion context \( (F (4, 7161) = 109.87, p < .001) \), but no main effect was found for talker group. Interactions between condition by emotion context \( (F (4, 7161) = 37.361, p < .001) \), talker group by emotion context \( (F (4, 7161) = 16.494, p < .001) \), and the three-way interaction between condition, talker group and emotion context \( (F (4, 7161) = 13.107, p < .001) \) were also significant. Overall, ERA for the f0-unmodified condition was higher (.52) compared to the f0-
modified condition (.46), and no difference in ERA between talker groups was found in either condition. However, taking a more careful look the three-way interaction between condition, talker group, and emotion context revealed that listeners did respond differently to stimuli from the two talker groups and the patterns of results was different for each emotion context.

Figure 4.2 displays ERA by condition, talker group and emotion context. Planned comparisons revealed that listeners more accurately identified phrases produced by talkers with ASD in the angry and sad contexts compared to those produced by TD talkers, but only in the f0-unmodified condition. Conversely, listeners had higher ERA scores for phrases produced by TD talkers in the happy and neutral contexts only in the f0-unmodified condition. No talker group differences were found for f0-modified stimuli in the angry, happy, neutral, and sad conditions. The interested context showed an opposite pattern – phrases produced by TD talkers in the interested context were identified at a higher rate compared to those produced by talkers with ASD, but only in the f0-modified condition. ERA decreased in the f0-modified condition for both talker groups in the angry, happy and interested emotion contexts, but increased for the neutral and sad contexts. Overall, this pattern of results shows that removing f0 variability in the recordings produced by talkers with ASD made the emotional content in the phrases sound more like the set of TD recordings.

The analysis of listener naturalness ratings revealed main effects for talker group ($F(1, 7161) = 31.567, p < .001$), condition ($F(1, 7161) = 257.236, p < .001$) and emotion context ($F(4, 7161) = 92.232, p < .001$), and significant interactions were found for talker group and condition ($F(1, 7161) = 21.869, p < .001$), emotion context and condition ($F(4, 7161) = 8.736, p < .001$), emotion context and talker group ($F(4, 7161) = 10.125, p < .001$), and the three-way
interaction between talker group, condition, and emotion context \( (F(4, 7161) = 5.552, p < .001) \). Overall, phrases produced by talkers with ASD received slightly lower naturalness ratings \( (5.61) \) than those produced by TD talkers \( (5.85) \), although a portion of the variability in the group means is subsumed by contrasting patterns of naturalness for the different emotion contexts.

![Figure 4.2](image)

Figure 4.2. Listener emotion recognition accuracy (ERA) for evoked phrases produced by talkers with ASD and TD controls. Points represent mean proportion correct scores for each emotion context (collapsed across repetition), talker group, and condition (modified or unmodified stimuli). Error bars represent +/- one 95% confidence interval around the mean.

Figure 4.3 displays mean listener naturalness ratings for each condition, talker group and emotion context. The f0-modified condition resulted in a general trend for listeners to rate the stimuli as sounding less natural than the f0-unmodified stimuli. Bonferroni-corrected post-hoc tests were performed to compare listener judgments of naturalness by condition, talker group, and context. In the f0-unmodified condition, listeners rated the happy and sad contexts as
sounding more natural; in the f0-modified condition, listeners rated angry, happy, and interested phrases produced by talkers ASD as less natural sounding compared to those produced by TD talkers.

Figure 4.3. Mean listener naturalness ratings for evoked phrases produced by talkers with ASD and TD controls. Points represent means for each emotion context (collapsed across repetition), talker group, and condition (modified or unmodified stimuli). Error bars represent +/- one 95% confidence interval around the mean.

4.2.7 Discussion

To investigate the impact of f0 variability on perception of affective prosody in ASD, a listening test was conducted in which modified and unmodified versions of the expressive phrases were presented to listeners, who were asked to identify the emotion context and rate the level of naturalness in each phrase. Emotion labels and naturalness ratings were obtained from a
group of TD listeners using a five-item forced-alternative listening task, and a slider-type scale for naturalness ratings. The overall pattern of responses for unmodified stimuli was that angry, interested, and sad phrases produced by talkers with ASD were identified at a higher rate than those produced by TD talkers, whereas the opposite pattern was found in the neutral and happy contexts. For both talker groups, presenting the f0-modified stimuli led to higher identification rates for neutral and sad contexts, and lower identification rates for angry, happy, and interested phrases.

Compared to the f0-unmodified phrases, presenting f0-modified phrases to listeners resulted in more similar ERA scores for the two talker groups. In four out of five emotion contexts (excluding interested), significant talker group differences found in the unmodified condition diminished in the modified condition. This result suggests that reducing f0 variability and modifying the f0 contours to match the mean for the TD group had a significant impact on listener perception, making the phrases produced by talkers with ASD sound more similar to those produced by TD talkers.

Concerning naturalness, there was a general trend for modified stimuli to sound less natural to listeners. For the unmodified condition, naturalness ratings were indifferent in three of the five emotion contexts, but happy phrases produced by TD talkers and sad phrases produced by talkers with ASD sounded more natural to listeners. In the modified condition, phrases produced by talkers with ASD were judged to sound less natural in the angry, happy, and interested contexts. This result suggests that modifying the f0 contour to match the mean for the TD group had a larger impact on naturalness ratings for phrases produced by talkers with ASD compared to those produced by the TD group.
The results of this experiment suggest that f0 variability is an important feature involved in perception of affective prosody, and that inconsistent patterns of f0 production are at least partly responsible for differences in listener perception. Abnormal prosody production in ASD is associated with negative social judgments by listeners (Shriberg et al., 2001; Van Bourgondien & Woods, 1992), and a likely contributor to negative judgments and perceptions of oddness is exaggerated f0 production. The finding that recordings produced by talkers with ASD were rated with lower naturalness is consistent with that conclusion. For individuals with ASD, a promising outcome is that enhancements in objective diagnostic methods and potential training methods could be developed based on measurement and reduction of f0 variability in affective prosody production.

4.2.8 Limitations

A potential limitation of this study is that only male talkers were included in the expressive speech database and listening test stimuli. Female talkers were excluded from the expressive speech database for several reasons. First, due to anatomical and social factors, male and female f0 measurements fall into different ranges. Male f0 measurements typically range from 85 to 180 Hz compared to 165 to 255 Hz for females. In addition, females have been shown to be more emotionally expressive compared to males (Brody & Hall, 1993; Kring & Gordon, 1998). Thus, as a first step, females were excluded to eliminate perceptual impacts from the use of different f0 registers and sex-based differences in production of affective prosody. Very little research has been conducted involving resynthesis and perceptual testing using recordings produced by adults with ASD, and as a first step males were included to establish a normative baseline for talkers with ASD and TD controls.
Another potential limitation involves the use of expressive speech portrayals produced in a laboratory setting. It is recognized that evoked expressive speech portrayals are not the same as normative production of expressive speech in an everyday setting. By definition, expressive speech that occurs in the presence of an experimenter is not naturally occurring. There is a tradeoff between spontaneity and experimental control in emotion research (Banziger & Scherer, 2007). A key objective for this project was to operate under a high degree of experimental control over the recording conditions, speech materials and elicitation methods to allow a comprehensive acoustic analysis of affective prosody and to generate high quality recordings to be used as test stimuli in emotion identification tasks. Here, the use of a controlled set of speech materials embedded in a role-playing task in which all talkers heard the same instructions is seen as an advantage over previously reported methods involving spontaneous and/or semi-spontaneous task. Interpreting the results of acoustic analyses of recordings in which the talkers produce different lexical content is difficult due to potential confounds between the speech content and production of affective prosody.

4.3 Experiment 2: The role of intensity in affective prosody perception

4.3.1 Introduction

Intensity is a measure of energy in the voice and is known to be an important acoustic property involved in communication of affective prosody. In the acoustic analysis presented in Section 3.5, talkers with ASD produced evoked expressive phrases with greater intensity (dB SPL) and intensity variability (SD in dB SPL) compared to TD talkers (see Table 3.1). This experiment was designed to study the perceptual impact of talker group-based intensity
differences on listener perception of affective prosody. Evoked expressive phrases were presented to listeners in two conditions consisting of unmodified and dB-modified versions of the original recordings. Evoked phrases were chosen due to the realistic nature of the recordings compared to nonsense syllables, and because the lexically-neutral phrases isolate perceptual patterns due to the talker’s acoustic modulations from those related to the verbal content of the phrases.

Prior research has shown that intensity increases in high-arousal emotion contexts (e.g., angry, happy) and decreases in low-arousal contexts (e.g., sad, neutral) relative to the mean (Scherer, 1979, 1986; Murray and Arnott, 1993; for a review see Juslin and Laukka, 2003). As one might imagine, the vocal effort required to produce the high arousal emotions is greater than that required to produce the low arousal emotions. The notable result presented in Section 3.5 is that talkers with ASD produced speech in non-neutral emotion contexts with greater mean intensity compared to TD talkers. Thus, a subset of the recordings was presented to listeners to test the impact of intensity differences and modifications to the intensity contour. To the extent that variation in intensity and greater overall intensity impacts perception of affective prosody, emotion identification for recordings produced by talkers with ASD will vary from those produced by TD talkers.

### 4.3.2 Method and procedure

A subset of evoked phrases selected from the expressive speech database were modified and presented to listeners in an emotion identification task. The same subset of phrases used in Experiment 1 served as test stimuli, however, the intensity contour of each phrase was modified instead of f0. Resynthesized versions of the database recordings were created by scaling the dB
SPL values over the course of each phrase to the mean for the TD group. This was accomplished using an overlap-add technique that involved splitting each test phrase into 1 ms segments, measuring the intensity value on a frame-by-frame basis, scaling that value within each frame to the mean value for the TD group, and synthesizing the dB-modified signal. Recordings for each evoked phrase and emotion context were peak-aligned prior to processing so that relevant features of each intensity contour were as aligned as possible prior to modifying the intensity values.

Listeners were presented with an equal number of dB-modified and dB-unmodified stimuli. During testing, an audio presentation of the test phrase was followed by five buttons presented on a computer screen to capture the emotion judgment for each stimulus. The button labels corresponded to the five recorded contexts discussed in Chapter 2: “neutral”; “angry”; “happy”; “interested”; and “sad”. The emotion buttons appeared on the computer screen in random order (per talker) to eliminate the possibility that the position of a button on the screen would impact the overall pattern of responses. Prior to beginning the emotion identification test, listeners were instructed to “listen carefully to each stimulus and identify which emotion context you heard by clicking the appropriate button on the computer screen” and that “some of the emotions are difficult to distinguish, but please click on the button that most closely matches the emotion you heard.”

Naturalness ratings were obtained using a slider-type scale ranging from 0 (least natural sounding) to 10 (most natural sounding). Listeners were given the following instructions: “after you select an emotion context the system will prompt you to rate how natural the phrase you heard was. For the purposes of this experiment, naturalness is defined as the degree to which the
emotional aspects of the phrase you heard would be encountered in everyday expressive communication. The naturalness rating is on a slider-type scale from 0 to 10 and your task will be to use the computer mouse to slide the marker to indicate a low naturalness rating on the left end of the slider or a high naturalness rating on the right side, or somewhere in between.” A 12-item practice session preceded the test and included stimuli that were not in the main experiment. For additional details about the listening test method and procedure, please see Section 4.1.

4.3.3 Participants

Twenty-one TD listeners served as participants in the listening test (16 females, 5 male, mean age 20 years). All listeners were native speakers of American English with no reported speech or hearing problems, were recruited from the UTD Behavioral and Brain Sciences participant pool and received research participation credit as compensation. Each listener self-reported no history of being diagnosed with autism, and completed the Broad Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007) to assess their tendency to display traits consistent the Broad Autism Phenotype (BAP). All participants had an overall BAPQ score below the increased specificity cutoff (3.55) suggested by Sasson and colleagues (2013), which suggests they did not display traits consistent with BAP.

4.3.4 Data analysis

Each listener provided emotion category judgments and naturalness ratings for 360 phrases (72 per emotion context, divided evenly by experimental condition). Emotion recognition accuracy (ERA) was calculated by dividing the number of identified phrases by the number produced in that context. Mean naturalness ratings were calculated by averaging
individual scores across items. The data were then summarized to enable comparisons by experimental condition, talker group and emotion context. ERA and naturalness were treated as dependent variables in separate mixed effects regression analyses, and talker group, emotion context, phrase and repetition were used as independent variables. Participants were treated as a repeated random variable given that each listener provided multiple emotion context judgments and naturalness ratings. The R statistical computing software (R Core Team, 2015) and the “lme4” package (Bates et al., 2015) were used to perform the ERA and naturalness analyses.

4.3.5 Hypotheses

Based on the acoustic analysis results presented in Section 3.5, it was predicted that ERA and naturalness ratings would be lower for the ASD talker group compared to the TD group, but only for the unmodified condition. Phrases produced by talkers with ASD were produced with greater mean intensity and intensity variability than phrases from TD talkers, and those acoustic differences were expected to negatively impact accuracy in the emotion identification task, as well as lower naturalness ratings. To date, very little research has been conducted involving voice intensity and perception of affective prosody in ASD; the primary aim of this study is to measure the impact of abnormal production of intensity on listener perception.

4.3.6 Results

For ERA, main effects were found for talker group \((F(1, 5366) = 11.320, p < .001)\) and emotion context \((F(4, 5366) = 71.236, p < .001)\). Significant interactions were found between condition and talker group \((F(1, 5366) = 3.984, p < .05)\), and talker group and emotion context \((F(4, 5366) = 34.125, p < .001)\). Overall accuracy for the dB-modified and dB-unmodified
conditions was indifferent (.55 vs. .54, respectively), however, divergent patterns of listener responses were revealed when the data were separated by condition and emotion context. The emotion context of unmodified phrases produced by talkers with ASD was correctly identified in 57% of trials, compared to 50% for unmodified phrases produced by TD talkers. In contrast, ERA for modified stimuli was the same for the two talker groups (56% for talkers with ASD, 54% for TD talkers).

Figure 4.4. 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 condition (dB-modified or dB-unmodified stimuli). Error bars represent +/- one 95% confidence interval around the mean.

Figure 4.4 displays ERA by condition, talker group and emotion context. The figure displays the emotion-specific group differences in ERA, and the interaction between talker group and condition. Bonferroni-corrected post-hoc tests were performed to further investigate the
significant interactions. For the condition by talker group interaction, a group difference in ERA was found only in the unmodified condition. There was a general trend toward reduced talker-group differences in ERA in the dB-modified condition. This trend reached significance in the interested context, where the talker-group difference in ERA diminished in the dB-modified condition.

Figure 4.5. Mean listener naturalness ratings for evoked phrases produced by talkers with ASD and TD controls. Points represent means for each emotion context, talker group, and condition (modified or unmodified stimuli). Error bars represent +/- one 95% confidence interval around the mean.

The analysis of listener naturalness ratings revealed a main effect for emotion context ($F(4, 5336) = 153.261, p < .001$) and a significant interaction for talker group and emotion context ($F(4, 5366) = 12.694, p < .001$). Overall, no difference was found between naturalness ratings for stimuli in the dB-modified (5.93) versus dB-unmodified (5.90) conditions, or between
talkers with ASD (5.89) versus TD talkers (5.94). Figure 4.5 displays mean listener naturalness ratings for each condition, talker group and emotion context. The figure shows the main effect found for emotion context – neutral stimuli sounded less natural to listeners compared to the other four emotion contexts. Bonferroni-corrected post-hoc tests were conducted to learn more about the interaction between talker group and emotion context. Mean naturalness ratings between talker groups significantly differed for the happy and sad emotion contexts. Listeners rated TD happy phrases as more natural sounding than happy phrases produced by talkers with ASD (6.59 vs. 5.89). The opposite pattern was true for sad phrases – those produced by talkers with ASD had a higher mean naturalness rating (5.96) compared to TD phrases (5.58).

4.3.7 Discussion

Experiment 2 was designed to examine the perceptual impact of increased variability in patterns of intensity production in evoked expressive phrases produced by talkers with ASD. The two objectives of the experiment were to determine if perceptual differences exist between phrases produced by talkers with ASD and controls, and to determine whether any group differences found are reduced or eliminated when the intensity contour is altered. Intensity-modified versions of expressive phrases were created by scaling the intensity contour to the mean for the TD talker group. Modified and unmodified versions of the phrases were presented to a group of TD listeners, and emotion category judgments and naturalness ratings were collected.

Emotion-specific talker-group differences were found in ERA and naturalness ratings. Angry, interested, and sad phrases produced by talkers with ASD were identified at a higher rate than those produced by TD talkers, yet ERA was generally higher for TD talkers in the happy
and neutral contexts. No differences in ERA or naturalness were found between the two experimental conditions (dB-unmodified vs. dB-modified). There was an overall trend for ERA to be more similar between talker groups in the modified condition compared to the unmodified condition in each emotion context except for neutral, which suggests that the group-based differences in affective prosody production diminished somewhat in the modified stimuli.

In general, however, scaling the intensity contours of to the mean for the TD talker group resulted in relatively small changes in ERA, as evidenced by the absence of a main effect of condition, and based on the result that the interaction between conditions and talker groups was significant only in the interested context. Additionally, relatively small changes in naturalness were found as a result of the dB modification. Naturalness ratings were different between groups only in the happy context, where phrases produced by talkers with ASD received lower naturalness ratings for both conditions, and in the sad context, where a group difference was found only in the modified condition.

In comparison with Experiment 1, the pattern of results for dB-unmodified stimuli matched that of f0-unmodified stimuli: for both experiments, listener ERA was lower for unmodified stimuli produced by talkers with ASD in the neutral and happy emotion contexts, and higher for talkers with ASD in the angry, interested, and sad emotion contexts. However, the relatively small changes in ERA as a result of dB-modification (and shallow slopes lines connecting experimental conditions in Figure 4.4) contrast the steeper slopes shown in Figure 4.2, particularly in the high arousal emotion contexts. Comparison of those results suggests that modifying the f0 contour had a larger impact on listener perception than changes to the intensity contour.
4.3.8 Limitations

As in Experiment 1, a potential limitation of this study is that only male talkers were included in the listening test stimuli. Female talkers were excluded from the expressive speech database to remove perceptual impacts due to anatomical and social factors leading to different f0 ranges for males and females, and potential sex-based differences in the use of affective prosody. Another potential limitation involves the use of expressive speech portrayals produced in a laboratory setting. It is recognized that evoked expressive speech portrayals are not the same as normative production of expressive speech in an everyday setting. By definition, expressive speech that occurs in the presence of an experimenter is not naturally occurring. There is a tradeoff between spontaneity and experimental control in emotion research (Banziger & Scherer, 2007). A key objective for this project was to operate under a high degree of experimental control over the recording conditions, speech materials and elicitation methods to allow a comprehensive acoustic analysis of affective prosody and to generate high quality recordings to be used as test stimuli in emotion identification tasks. Here, the use of a controlled set of speech materials embedded in a role-playing task in which all talkers heard the same instructions is seen as an advantage over previously reported methods involving spontaneous and/or semi-spontaneous elicitation methods.

4.4 Experiment 3: Affective prosody perception in listeners with ASD

4.4.1 Introduction

The third experiment was designed to investigate perception of affective prosody in listeners with ASD. Compared to studies of affective prosody production, less attention has been
devoted to perception differences in ASD listeners. To date, no known studies have examined affective prosody perception in listeners with ASD using recordings produced by both talkers with ASD and TD controls. In the current experiment, expressive phrases from by each talker group produced in five emotion contexts were presented to listeners, and emotion labels and naturalness ratings were collected. The design of the experiment allowed the ability to examine listener group-based perception differences as well as interactions between talker and listener groups.

Research in this area has revealed affective prosody perception deficits for children with ASD. In Peppé et al. (2007), expressive and receptive prosodic abilities in ASD were examined using the Profiling Elements of Prosodic Systems in Children test (PEPS-C; Peppé & McCann, 2003). The study included a large sample of children from 6 to 13 years of age with ASD and TD controls (ASD: 31; TD: 72) and a comparison group of 33 adults. The affect reception component of the test is a picture naming task in which participants hear one of four words (tea, mushrooms, milk, cream) spoken in a positive tone to indicate liking or a negative tone to indicate disliking. Participants click on a happy face displayed on a computer screen to indicate liking or a sad face to indicate disliking. Results revealed significant differences between the three groups on the affect reception task. The adult comparison group had the highest mean score of the three groups (96.5; $SD = 4.9$), followed by TD children (84.5; $SD = 11.4$) and children with ASD (71.2; $SD = 21.6$). An interesting result was that score variability ($SD$) was negatively correlated with performance – children with ASD had the greatest score variability and the lowest performance on the test.
Additional work has shown that listeners with ASD have difficulty detecting vocal cues for complex emotions such as irony and sarcasm. In a study that presented cartoon drawings and audio vignettes of children in conversational settings, listeners with ASD identified irony with lower accuracy compared to TD listeners (Wang et al., 2007). Another study involved a commonly used set of recordings designed to assess listeners’ ability to extract emotional content from vocalizations known as the “Reading the Mind in the Voice” test (Rutherford et al., 2002). The revised version (RMV-R) was administered to adults with ASD as well as TD control listeners, resulting in significant differences between listener groups (Golan et al., 2007). Listeners with ASD had difficulty recognizing complex emotions from test stimuli.

In the current study, expressive phrases produced in five emotion contexts were presented to listeners with ASD and TD controls to examine perception of affective prosody in the two listener groups. In comparison with the studies described above, the current study involved a 2x2 experimental design in which ASD and TD listener groups responded to phrases produced by ASD and TD talker groups. Therefore it was possible to test for interactions between the talker and listener groups and determine whether participants from each listener group show affective prosody perception differences and whether those differences are linked to inconsistent patterns of production.

4.4.2 Method and procedure

The subset of the evoked phrases presented to listeners in Experiment 3 included 330 recordings. An equal number of stimuli were presented from each talker group and emotion context. Each of the 330 trials consisted of listening to a test stimulus, judging the emotion context, and then rating the level of naturalness of the emotional content in the phrase. All phrase
types and talkers that were recorded as part of the expressive speech database were included in
the set of test stimuli (see Chapter 2). Listening tests were performed inside a sound-attenuated
booth while seated listening to phrases over headphones. For additional details about the
listening test method and procedure, please see Section 4.1.

4.4.3 Participants

Thirty TD listeners (20 females, 10 males, mean age 22.5 years) and 22 listeners with
ASD (2 females, 20 males, mean age 25.9 years) served as participants. All listeners reported
being native speakers of American English with no speech or hearing problems. Participants with
ASD were recruited through the UTD Autism Research Collaborative (ARC), a database of
research participants with a diagnosis of ASD. Diagnoses of ASD were confirmed by a certified
clinician based on the Autism Diagnostic Observation Schedule (Lord, et al., 2000). Participants
with ASD were provided with compensation for producing the set of recordings.

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 test for the Broad Autism Phenotype (BAP). Each
TD listener self-reported no history of being diagnosed with autism and completed the Broad
Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007) to assess their tendency to display
traits consistent the Broad Autism Phenotype (BAP). Three listeners had an overall BAPQ score
above the increased specificity score (3.55) for BAP suggested by Sasson and colleagues (2013);
the listening test results for those listeners were included in the analysis after individual-level
comparisons confirmed that the results closely matched the overall pattern.
4.4.4 Data analysis

Each listener provided emotion category judgments and naturalness ratings for 330 phrases (66 per emotion context, divided evenly by experimental condition). Emotion recognition accuracy (ERA) was calculated by dividing the number of identified phrases by the number produced in that context. Mean naturalness ratings were calculated by averaging individual scores across items. ERA and naturalness were treated as dependent variables in separate mixed effects regression analyses, and listener group, talker group, emotion context, phrase and repetition were used as independent variables. Listeners were treated as a repeated random variable given that each provided multiple emotion context judgments and naturalness ratings. The R statistical computing software (R Core Team, 2015) and the “lme4” package (Bates et al., 2015) were used to perform the ERA and naturalness analyses.

4.4.5 Hypotheses

An interesting pattern of results has emerged in the perception literature in that compared to TD listeners, ASD participants show increased proficiency in tasks involving low-level acoustic information such as discrimination of relative pitch using simple tones (Heaton, 2005; Bonnel, et al., 2010). In contrast, ASD participants show impaired performance in tasks involving more spectrally or temporally complex stimuli including speech (Whitehouse & Bishop, 2008; Kuhl et al., 2005). Based on the results described above, it was predicted that emotion recognition accuracy (ERA) would be significantly lower for listeners with ASD compared to TD listeners. While the emotions presented to listeners in this study (angry, happy, interested, sad, and neutral) would be considered more basic than irony and sarcasm (in contrast
with Wang et al., 2007), it seems likely that deficits in affective prosody carry over at least to some degree to the more basic emotions. That pattern would be expected if disordered affective prosody perception is a symptom of more general deficits in emotion processing in ASD.

4.4.6 Results

For ERA, main effects were found for talker group \((F(1, 17090) = 123.759, p < .001)\), listener group \((F(1, 50) = 7.064, p < .05)\), and emotion context \((F(4, 17090) = 127.877, p < .001)\). Significant interactions were found between listener group and emotion context \((F(4, 17090) = 9.427, p < .001)\) and talker group and emotion context \((F(4, 17090) = 66.606, p < .001)\). Overall, phrases produced by talkers with ASD were identified at a rate of 56%, compared to 48% for TD talkers, and the TD listener group achieved a higher overall ERA score (54%) than the group of listeners with ASD (49%).

Figure 4.6 displays ERA by talker group, listener group and emotion context. All mean ERA scores were well above chance level, which for the five-item forced-alternative task was 20%. The lowest mean ERA was for happy phrases produced by TD talkers (36%). Bonferroni-corrected post-hoc tests were performed to further examine the interactions between talker group and emotion context, and listener group and emotion context. Significant talker group differences were found for listeners with ASD in the neutral, angry, happy, and sad contexts, and for TD listeners in the neutral, angry, and sad contexts. With the exception of the neutral context, phrases produced by talkers with ASD were more accurately identified than those produced by TD talkers. The largest talker group differences were found in the angry and sad contexts, where ERA increased for talkers with ASD by approximately 20 percentage points. The reverse pattern 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. Regarding the listener
group/emotion context interaction, significant difference in ERA were found between listener
groups only for phrases produced in the neutral and happy contexts. For the neutral context,
phrases produced by talkers with ASD and controls were more accurately identified by TD
listeners than listeners with ASD. In the happy context, ERA was increased for phrases produced
by TD talkers, but only for TD listeners. Performance was indifferent for phrases produced by
talkers with ASD for both listener groups.

![Figure 4.6](image)

Figure 4.6. 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.

The analysis of listener naturalness ratings revealed main effects of talker group \( (F (1, 17090) = 98.911, p < .001) \) and emotion context \( (F (4, 17090) = 141.600, p < .001) \), and
interactions were found for the comparisons between talker group and listener group \((F(1, 17090) = 10.792, p < .01)\), listener group and emotion context \((F(4, 17090) = 34.618, p < .001)\), and talker group and emotion context \((F(4, 17090) = 11.136, p < .001)\). Overall, phrases produced by talkers with ASD received slightly lower naturalness ratings (5.74) than those produced by TD talkers (6.05), and angry phrases were rated as the most natural sounding (6.31) compared to neutral phrases, which were rated as the least natural (5.10). Figure 4.7 shows mean listener naturalness ratings for each talker group and emotion context.

Figure 4.7. 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.

Bonferroni-corrected post-hoc tests were performed to further explore the interactions between the talker and listener groups and emotion context. In the happy and interested contexts, phrases produced by TD talkers were rated more natural by both groups of listeners. There was
also a general trend for phrases heard by TD listeners to be rated less natural sounding compared to those heard by listeners with ASD, particularly for speech produced by TD talkers. This effect reached significance in the interested and sad contexts for TD talkers and in the neutral context for both talker groups.

### 4.4.7 Discussion

Perception of affective prosody was examined in listeners with ASD using an emotion identification test. Expressive phrases were presented to a group of listeners with ASD and a group of TD controls, and emotion context labels and ratings of naturalness were collected. In addition, half of the phrases presented as listening test stimuli were produced by talkers with ASD and half were produced by TD talkers. This method provided the ability to examine group-based effects in ERA and naturalness in the talker and listener groups, as well as interactions between the two group types. In general, talker-group effects in ERA and naturalness were larger than listener-group effects. The overall pattern of results showed that regardless of listener group. Compared to phrases produced by TD talkers, the emotion context of phrases produced by talkers with ASD was more likely to be identified, and naturalness ratings were more likely to be lower. A notable exception was found in the neutral context – neutral phrases produced by the TD talker group were more accurately identified than those produced by the group of participants with ASD.

The results did not support the prediction that listeners with ASD would exhibit lower ERA scores compared to controls. Instead, the results for listeners with ASD generally matched those for TD listeners. Listener-group differences were found only in the neutral context and in happy phrases produced by talkers with ASD (see Figure 4.4). Indeed there are mixed results in
the literature regarding affective prosody perception in ASD. Deficits in perception of affective receptive have been reported for participants with ASD in some studies (e.g., Peppé et al., 2007; Philip et al., 2010), while similar performance to controls has been reported in others (e.g., Boucher et al., 2000; Chevallier et al., 2011). The discrepancy in results is likely due to differences in study participants, stimulus type and complexity. O’Conner (2007) and Doyle-Thomas et al. (2013) reported equivalent performance in listeners with ASD and a control group when stimuli were presented in isolation, which is consistent with the results reported here, but the authors found impairments in emotion recognition in participants with ASD when expressive speech was presented together with emotional faces. This result suggests a relationship between task complexity and performance that was not examined here.

Key differences in design between the current study and previously published work include that prior research on perception of affective prosody in ASD has been conducted primarily using children (e.g., Peppé et al., 2007; ), and affect expression and affect reception tasks in the PEPS-C are quite different than the methods used in the current study. Here, instead of labeling the emotion context of single words, phrases of between 4 and 5 words were used, which could contribute to more similar ERA scores from listeners with ASD and controls given additional speech content to utilize in emotion labeling.

4.4.8 Limitations

As in Experiments 1 and 2, a potential limitation of this study is that only male talkers were included in the listening test stimuli. Female talkers were excluded to avoid potential confounds due to anatomical and social factors leading to differences in f0 ranges for males and females, and potential sex-based differences in the use of affective prosody. Another aspect that
may be seen as a limitation involves the use of expressive speech portrayals produced in a laboratory setting. It is recognized that evoked portrayals are not the same as normative production of expressive speech in an everyday setting, however, the fact that the overall ERA scores were well above chance for both listener groups suggests that the expressive speech portrayals are good exemplars of the intended emotion contexts. By definition, expressive speech that occurs in the presence of an experimenter is not naturally occurring, and there is a tradeoff between spontaneity and experimental control in emotion research (Banziger & Scherer, 2007).

A key objective for this project was to operate under a high degree of experimental control over the recording conditions, speech materials and elicitation methods to allow a comprehensive acoustic analysis of affective prosody and to generate high quality recordings to be used as test stimuli in emotion identification tasks. Here, the use of a controlled set of speech materials embedded in a role-playing task in which all talkers heard the same instructions is seen as an advantage over previously reported methods involving spontaneous and/or semi-spontaneous tasks.

### 4.5 Modeling perception of affective prosody

The production and perception data for evoked expressive phrases were fit with a multinomial mixed effects logistic regression model to predict listener responses based on the set of acoustic variables investigated in this project. Multinomial regression uses maximum likelihood estimation to predict category membership based on one or more independent variables (Hosmer, 2013). Here, the proportion of angry, happy, interested, sad, or neutral responses to the set listening test stimuli was modeled using talker group (ASD or TD), f0 range,
mean intensity, and mean phrase duration as independent variables. The R “mlogit” package (Croissant, 2013) was used to perform the multinomial logistic regression.

The overall model fit was significant, $\chi^2 (4, N=30) = 3075.5$, $p < .001$ and predicted listener responses were similar to the original response probabilities. Predicted emotion context responses were computed to examine the relationship between each of the acoustic variables and listener responses. Figures 4.8 through 4.12 show predicted response curves separately for each emotion context while varying f0 and intensity.

Figure 4.8. Predicted probabilities for selecting the neutral emotion context based on changes in f0 range (left panel) and intensity (right panel). Shaded regions represent 95% confidence bands around the predicted functions.

Figure 4.8 shows predicted neutral responses as a function of changes in f0 and intensity. For talkers with ASD, the probability of selecting neutral drops as f0 range rises above 60 Hz; for TD talkers the probability remains consistent at just under 0.5. The probability of selecting neutral drops for both talker groups as mean intensity increases. The probability of selecting neutral was higher for the TD talker group, independent of changes in f0 and intensity. Figure
4.9 displays predicted probability curves for selecting the angry context. The model predicted a higher proportion of angry responses for talkers with ASD compared to controls. The probability of selecting angry remains consistent (at around 0.3) for talkers with ASD and near chance level (0.2) for TD talkers despite changes in f0 range, but increases with greater mean intensity for both groups.

Figure 4.9. Predicted probabilities for selecting the angry emotion context based on changes in f0 range (left panel) and intensity (right panel). Shaded regions represent 95% confidence bands around the predicted functions.

Figure 4.10 displays predicted probabilities for selecting the happy emotion context for the ASD talker group and TD controls as a function of variation in f0 and intensity. The predicted likelihood of selecting happy increases dramatically for both talker groups as f0 range increases, and the model predicts similar results for the two talker groups. In fact, the model predicts below chance probabilities of selecting happy when f0 range is below 400 Hz. Conversely, mean intensity and predicted happy responses are negatively related, such that predicted happy responses decline sharply as intensity rises above 55 dB SPL.
Figure 4.10. Predicted probabilities for selecting the happy emotion context based on changes in f0 range (left panel) and intensity (right panel). Shaded regions represent 95% confidence bands around the predicted functions.

Figure 4.11. Predicted probabilities for selecting the interested emotion context based on changes in f0 range (left panel) and intensity (right panel). Shaded regions represent 95% confidence bands around the predicted functions.
Figure 4.11 shows predicted probability curves for selecting the interested emotion context for the ASD talker group and TD controls. The probability of selecting interested is consistent for both groups as f0 range increases, but decreases for both talker groups as mean intensity increases. Overall, the model predicts that listeners will be more likely to select the interested context when the speech is produced by TD talkers, despite changes in f0 range or mean intensity. For the sad context, predicted response curves were similar for each group for changes in f0 and intensity (see Figure 4.12). The model predicted some fluctuation in sad responses due to variation in f0 range, particularly for the ASD talker group, but the overall response probabilities were similar for the two groups.

In summary, the model predictions were generally consistent with expected patterns of change in acoustic properties relevant for affective prosody production, and the predicted
emotion contexts in many cases matched the actual responses from TD listeners. The results highlight the fact that affective prosody communication involves expression and reception of multiple voice properties used as emotion cues in expressive communication. Figures 4.8 through 4.12 reveal complex relationships between predicted emotion category responses as a function of variations in f0 and intensity, and distinct patterns for the two talker groups. In some cases (e.g., happy), f0 seemed to be more important for predicting emotion context responses, as indicated by steeper predicted response curves, while intensity seemed to be more important for others (e.g., neutral). For individuals with ASD, the model predicts that abnormal production of affective prosody leads to differences in listener perception that could provide insight into reasons talkers with ASD are sometimes perceived as socially odd. It is only through a deeper understanding of the use of affective cues in speech that we may one day be able to reduce impressions of oddness in the expressive speech of talkers with ASD.
CHAPTER 5
GENERAL DISCUSSION AND CONCLUSION

The research described in this thesis examined affective prosody functioning in ASD by conducting a series of experiments involving production of expressive speech recordings and perceptual testing using those recordings. The major findings were that adult males with ASD produced affective prosody that in many cases revealed inconsistent group patterns with more extreme emotional content compared to the TD control group, as indicated by the analysis of f0 range, intensity, and duration. In general, talkers with ASD produced the evoked phrases and VCV syllables in non-neutral emotion contexts with increased f0 range, greater mean intensity, and in some cases longer phrase and syllable durations compared to TD talkers.

The affective prosody production differences can be summarized by the outcome that talkers with ASD displayed an overall pattern of more exaggerated forms of affective speech production relative to TD talkers. This pattern of results held up for the evoked phrases and VCV syllables. In the scripted dialogue recordings, however, intensity and duration were equivalent in the two talker groups, but f0 range was still significantly higher for talkers with ASD. Talkers may have been able to rely on lexical information in the context-specific scripted dialogues that was not available in the evoked phrases and VCV syllables.

When a subset of the evoked phrases was presented to listeners, phrases produced by talkers with ASD were generally labeled with higher identification rates but lower naturalness ratings compared to phrases produced by TD talkers. This finding suggests that group differences
in affective prosody production impact listener perception, and in this case, led to increased emotion context identification and lower ratings of naturalness for talkers with ASD. Additional support for this conclusion was found in the results of Experiment 1 (see Section 4.2), which showed that removing f0 variability in the phases by replacing the f0 contour with the mean f0 contour in productions by the control group led to more similar perceptual results between the two groups compared to the unmodified condition.

The results suggest that f0 may be a more salient voice cue than intensity in perception of affective prosody. Comparing the results for Experiments 1 and 2 shows a relatively smaller perceptual impact for the intensity modification in Experiment 2 compared to f0 modification in Experiment 1. Scaling the intensity contours of test stimuli to match the mean for the TD group resulted in only small changes in ERA and naturalness. ERA in the f0-modified condition was equivalent for the two talker groups in four out of five emotion contexts due to the reduction in f0 variability in Experiment 1, while ERA was equivalent in only the interested context in dB-modified test stimuli from Experiment 2.

Several hypotheses have been offered to explain divergent patterns of prosody production in ASD. One potential explanation which has gained recent research interest is referred to as the social motivation theory (Chevallier et al., 2012). Research has shown that individuals with ASD are less likely to pay attention to salient social stimuli compared to control subjects (Klin et al., 2003). The focus on motivational factors represents a shift from the more traditional account that social impairments in ASD are a result of deficits in social cognition (e.g., “theory of mind”; Baron-Cohen, Leslie, & Frith, 1985). Rather, the social motivation theory suggests that problems with social cognition are a result of reduced motivation to attend to socially-relevant stimuli, thus
depriving developing children of important learning opportunities. Inattention to social norms could contribute to divergent patterns of affective prosody production, in fact, greater variability during production of acoustic properties relevant to affective prosody would be predicted by reduced learning about social norms for prosody.

A logical extension to the diminished social motivation account is that prosody functioning in individuals with ASD will be enhanced if attention is heightened by the study procedures. This was supported in an fMRI study involving perception of irony in speech (Wang et al., 2007). Wang and colleagues varied the type of instructions that listeners were exposed to, ranging from “pay close attention” to “pay close attention to the face and voice.” The authors reported increased activity in the medial prefrontal cortex in ASD listeners, an area linked to perception of irony in TD listeners, when they heard more explicit instructions. Additional support for this idea comes from face perception studies which have found that the accuracy of identifying emotional content in faces is dependent on task demands (Begeer et al., 2006; Kahana-Kalman & Goldman, 2008).

However, the social motivation theory does not account for the current result that overall listeners with ASD performed as well as TD listeners in a perceptual test involving emotion identification. Relative to listeners with ASD, higher ERA was found for TD listeners only in the neutral context and for happy phrases produced by TD talkers. In addition to expressive differences, the social motivation theory would predict inconsistent patterns of affective prosody perception as well, due to inattention to socially relevant affective cues in the phrases, which was not shown here. The perception results in Experiment 3 were largely attributable to the talker group the stimuli belonged to, rather than the listener group responding.
An alternative explanation for exaggerated patterns of affective prosody production involves speech attunement, which suggests that individuals with ASD have difficulty emulating models provided by other talkers due to challenges in social reciprocity and communication intent (Shriberg et al., 2011). The speech attunement framework is based on the validity of two assumptions. First, many individuals with ASD show enhanced pitch perception compared to controls when the stimuli consist of simple tones (Samson et al., 2006; Bonnel et al., 2010), reflecting a heightened ability to attend to basic acoustic properties. In addition, there is a higher prevalence of absolute pitch in individuals with ASD, which is the ability to recognize or reproduce any given tone without a reference tone (Miller, 1999; Heaton, 2005). Second, individuals with ASD may not experience strong motivation to attune their speech to match TD productions. This leads to a lack of social motivation to sound “just like” other talkers (Paul et al., 2008; Shriberg et al., 2011; Diehl & Paul, 2012), labeled by Shriberg et al. (2011) as speech attunement, referring to difficulties in “tuning up” speech to match socially acceptable standards.

The speech attunement framework is consistent with the current results showing divergent patterns of production in f0, intensity and duration, and could fit with the finding that listeners with ASD performed similarly to the control group in perceptual testing. A number of studies have reported decreased ability to identify emotions in speech by listeners with ASD (e.g., Paul et al., 2005; Peppé et al., 2007; Stewart et al., 2012); however, no known studies have included speech produced by talkers with ASD as listening test stimuli.

The current result that in most emotion contexts ERA was equivalent for listeners with ASD and controls, together with the finding that identification rates were higher for non-neutral phrases produced by talkers with ASD, suggests that exaggerated affective prosody production in
the recordings from talkers with ASD was interpreted in the same way by both listener groups. If deficits in reception of affective prosody exist in individuals with ASD, it may be that those deficits were overshadowed in the current project by exaggerated and comparatively more inconsistent versions of the expressive speech recordings produced by talkers with ASD.

Future research in this area should further investigate production and perception of affective prosody using a variety of theoretically-driven tasks and levels of stimulus complexity. For example, given that research has revealed different outcomes for talkers and listeners with ASD as a function of type of stimulus and stimulus complexity, it is suspected that longer segments of speech will be correlated with increased ERA, and that phrase length will likely interact with lexical ambiguity. Thus, a possible next step would be to design an experiment that systematically varies both of those variables in the set of test stimuli. It is only through a deeper understanding of the complex nature of prosodic deficits in ASD in terms of production, perception, and the interaction between the two, that we will be able to enhance diagnostic and training tools that may provide great benefit for individuals with ASD. In closing, it is important to acknowledge the heterogeneity and severity of symptoms in ASD. The talkers and listeners with ASD who participated in this study could be classified as high-functioning, and more research is needed to better understand the relationship between levels of functioning in ASD and social communication skills involving speech and emotion.
REFERENCES


VITA

Daniel J. Hubbard was born in Dallas, Texas in 1977 as the second of three sons to Ron and Linda Hubbard. In 2001, he graduated from The University of Texas at Dallas with Summa Cum Laude honors with a Bachelor’s Degree in Psychology. His honors thesis was entitled “Student retention efforts: Programs that work,” a project that detailed various types of programs designed to increase student success. As part of the internship program during his senior year at UT Dallas, Daniel began working as an intern for El Centro College of the Dallas County Community College District (DCCCD) as a research assistant in the Office of Institutional Effectiveness and Research. The internship led to full-time employment in progressive roles at El Centro College as a research assistant, research associate, and coordinator of statistical processes. Daniel continued to work for DCCCD as Director of Institutional Research at Brookhaven College while completing his Master’s Degree in Applied Cognition and Neuroscience at UT Dallas in 2006. Following completion of the Master’s Degree, Daniel taught introductory courses in Psychology at Brookhaven College and was accepted into the doctoral program in Cognition and Neuroscience at UT Dallas in 2008. Working as a part-time student in the doctoral program, Daniel accepted a new role as Dean of Planning and Research at Cedar Valley College, another of the seven colleges of the DCCCD. In 2012, Daniel resigned from Cedar Valley College to devote full-time effort into obtaining his Ph.D.