Prosody of Spontaneous Speech in Autism

Géza Kiss
M. S. in Computer Science, Budapest University of Technology, 1997

Presented to the
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree
Doctor of Philosophy
in
Computer Science & Engineering

September 2017

Center for Spoken Language Understanding
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph. D. dissertation of

Géza Kiss

has been approved.

Jan P.H. van Santen, Thesis Advisor
Professor

Alison Presmanes Hill
Associate Professor

Xubo Song
Professor

Éric Fombonne
Professor

Elmar Nöth
Professor

Dedication

To Klári and Peti, who were with me through thick and thin.

To Father and Mother, who did all in their power to help me achieve this goal.

Acknowledgments

First of all, I would like to thank my advisor, Jan van Santen, who has been instrumental in making this PhD training come to fruition. He took on the role of mentoring me during my Fulbright scholarship and then for the entire duration of my PhD program. He made it possible for me to move back to my home country before finishing, taking on the inconveniences associated with being a remote advisor. He was kind, encouraging, constructive, and supportive all the time. He let me pursue my research interests, while also providing goals, advice, and feedback. Last but not least, he was also the one who found the financial means for this undertaking.

I also want to express my sincere thanks to the professors who were on my advisory committee during these years, namely Alison Presmanes Hill, Esther Klabbers, and Xubo Song, and my external committee members, Éric Fombonne and Elmar Nöth. Their feedback made a vast difference in improving the quality of this dissertation. Kyle Gorman, Eric Morley, and Emily Tucker Prud'hommeaux also provided me with helpful feedback on parts of the manuscript.

I am grateful for the guidance of Peter Heeman through the PhD process, and the constant and apt support with administrative issues and other everyday matters from Patricia Dickerson. Steven Bedrick, Jason Brooks, Sean Farrell, Ethan Van Matre, Izhak Shafran, and Robert Stites always helped with issues related to the CSLU computer cluster. I am also grateful to Kim Basney for calling my attention to a tax treaty, which step improved our financial standing substantially.

The classes with professors like Steven Bedrick, John-Paul Hosom, Alexander Kain, Brian Roark, Izhak Shafran, and Richard Sproat, the weekly seminars, and the interaction with fellow students all made CSLU a fun and inspriring place to be. Thank you all for sharing some of the good treasures of your heart! It was great to be among such clever and friendly people.

I want to thank Pastor Fred and Betty Cason of True Life Fellowship in Beaverton, who were friends to us and helpful in all possible ways during our stay in the US, and provided a home for me during my later visits.

The word gratitude is not enough to express what I feel toward my wife, Klári, who loved and supported me faithfully me even when it was not so easy, and my oldest son, Peti, who was with us through it all. Above all, I thank Jesus Christ, without whom this dissertation would never have been completed.

This material is based upon work supported in part by the National Science Foundation under Grant No. NSF IIS-0905095 (van Santen, PI), and the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health under award numbers NIH-1R01DC007129 (van Santen, PI), NIH-1R01DC012033 (Richard Sproat, PI and van Santen, PI), and NIH-R21DC010239 (Lois Black, PI and van Santen, PI). The content is solely the responsibility of the author and does not necessarily represent the official views of the National Science Foundation or the National Institutes of Health.

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Expansion
ADOS	Autism Diagnostic Observation Schedule; see Section 4.2.1
ALI	Autism with Language Impairment
ALN	Autism with Normal Language
AMT	Amazon Mechanical Turk (a crowdsourcing website)
AS	Asperger's Syndrome
ASD	Autism Spectrum Disorders
ASR	Automatic Speech Recognition
ATT	Average Treatment Effect for the Treated
BEC	Best Estimate Clinical consensus judgment
CA	Chronological Age
CELF	Clinical Evaluation of Language Fundamentals
CoV	Coefficient of Variation
CSLU	Center for Spoken Language Understanding
DSM	Diagnostic and Statistical Manual of Mental Disorders
DX	diagnosis
F0	Fundamental frequency (pitch)
FDA	Functional Data Analysis
fMRI	Functional Magnetic Resonance Imaging
fPCA	Functional Principal Component Analysis
HFA	High-Functioning Autism
HIT	Human Intelligence Task (an AMT concept)
ID	Intellectual Disability
IQ	Intelligence Quotient
IQR	Inter-Quartile Range
L1LM	Linear Model with L1 regularization

Abbreviation	Expansion
LD	Learning Disabilities
LDA	Linear Discriminant Analysis
LI	Language Impairment
LMA	Lexical Mental Age
MAD	Median Absolute Deviation from the median
MELM	Mixed Effect Linear Model
MEP	Matched on Expected Prosody stimulus set; see Section 7.2.5
MID	Maximally Individually Diverse stimulus set; see Section 7.2.6
NVIQ	Non-Verbal IQ
OHSU	Oregon Health and Science University
PCA	Principal Component Analysis
PEPS-C	Profiling Elements of Prosodic Systems–Children; see Section 2.2.4
PIQ	Performance IQ (or Non-Verbal IQ)
POS	part-of-speech
PVSP	Prosody-Voice Screening Profile; see Section 2.2.3
RMS	Root Mean Squared value
RMSE%	Root Mean Squared Error relative to the range
SALT	Systematic Analysis of Language Transcripts; see 4.2.5
SD	Standard Deviation
SLAM	Simplified Linear Alignment Model
SLD	Spoken Language Disorders
SLI	Specific Language Impairment
SVR	Support Vector Regression
TD	Typical Development
VIQ	Verbal IQ
VMA	Verbal Mental Age

Abstract

Prosody of Spontaneous Speech in Autism

Géza Kiss

Doctor of Philosophy
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine

September 2017
Thesis Advisor: Jan P.H. van Santen

The goal of this work was to study speech prosody in Autism Spectrum Disorders (ASD). Increasing our understanding of how prosody is different in ASD may be important for characterizing its phenotype, helping prosodic remediation, aiding in diagnosis, and providing outcome measures for treatment research. We compared the prosody of children with High-Functioning Autism (HFA) to those with Specific Language Impairment (SLI) and Typical Development (TD). Our questions included the following: What statistical features of prosody are significantly different? How does intonation differ qualitatively, for example regarding the shape of the intonation curves? Can naive listeners reliably detect atypical prosody at the utterance level? After explaining the necessary scientific background, we first matched the groups on age and cognitive measures using a novel approach and new algorithms. Subsequently, we compared the prosody of these matched groups based on acoustic-prosodic features from various known and innovative computational techniques, as well as through perceptual ratings of the children's utterances from naive listeners. My main contribution to knowledge is that high-functioning autistic children without language impairment differed from typical children on various measures of prosody, whereas the autistic children with language impairment did not differ from children with specific language impairment on any of our measures after controlling for content features. The utterances of children with either HFA or SLI were also perceived as having higher emotional arousal that of those with TD.

Chapter 1

Introduction

Speech is different from the written word because it carries more than just the words: It has its own music, called the prosody of speech. This includes the melody or intonation, the rhythm, the variations of intensity, varying amounts of silence or pausing as well as the timbre of the voice — just like in a song. It depends on the particular words being spoken as well as the attitude and emotions of the speaker, and what he or she considers important in the message. By paying attention to the music of speech, we can even distinguish the person speaking, and his or her voice characteristics, just like we can distinguish the timbre of musical instruments, and perhaps even the identity of the musician based on his or her unique style.

But not all variation in the prosody of speech is due to differences in the words, the intention or even the personality of the speaker: Clinicians and researchers have regularly described prosodic differences compared to typical development in the speech of people with certain conditions, such as apraxia, Down's syndrome, Parkinson's disease, or autism. Their speech can sound atypical even when the message being spoken is grammatically and semantically correct so much so that it can be one of the most outstanding features of their condition.

Atypicality of a person's speech prosody is significant for multiple reasons. It may stem from a defect in the receptive abilities, that is, the person may not understand the use of prosody adequately, which hinders him or her not just in using it, but also in comprehending the speech of others. This may be one contributing factor to a lesser understanding of others, that is Theory of Mind (TOM) problems. But even if it is a problem with the expression only, it can hinder being understood, or the speaker may even be perceived as speaking inappropriately.

1.1 Problem Statement

The goal of this work is to study prosody in one particular condition, namely Autism Spectrum Disorders (ASD). Even the very first description of ASD by Kanner (Kanner, 1943) contains references to peculiarities in intonation, namely the verbatim repetition of the speech of others including the intonation, monotonous reading, humming, and singing, and sometimes odd intonation. Exactly why their prosody sounds atypical, whether there is a systematic difference that distinguishes it from typical speech, and if so, whether that difference distinguishes ASD from other conditions with a prosodic disorder, is not yet a settled question. Prosody has not been included in diagnostic procedures for ASD, at least in part due to reliability and validity issues: Individual judges are not consistent even within themselves, that is a person would often judge the same utterance differently on different occasions, and it is not clear how well a group of judges agree among each other.

Understanding how prosody is different in ASD is important for multiple reasons. First, understanding what has gone awry can help in remediation, that is, in training autistic people to express themselves in a more socially acceptable way and to understand others better, putting aside this barrier from social integration (Klin et al., 2007). Second, better characterization of the phenotype of ASD, and possibly the identification of subgroups (such as those with and without a prosodic disorder) is an important prerequisite of genetic research. Third, seeing what aspects of prosody are specific to autism, and what peculiarities occur in other disorders as well, may help in differential diagnosis, for example distinguishing ASD from language disorder or from Attention-Deficit Hyperactivity Disorder (ADHD). Fourth, it can contribute to the development of tools that may help in screening, and possibly help in the diagnostic procedure. Fifth, being able to assess prosody automatically can provide us with outcome measures for measuring the effectiveness of autism treatments.

1.2 Aims

We intend to contribute to the growing body of literature on prosodic differences in ASD compared to Typical Development (TD) and Specific Language Impairment (SLI) in childhood. Our aim was to collect and analyze data by applying newly developed or existing technologies with the intent of shedding light on questions that do not yet have a definite answer, such as:

- What acoustic features are significantly different in *High-Functioning Autism* (HFA, i.e., subjects with ASD who have an IQ of at least 80)?
- How is prosody in HFA different qualitatively, for example regarding the shape of the intonation curves?
- What is the relationship of this prosodic difference to the content? For example, is the prosody atypical given the words in the sentence, or atypical in itself and unlikely for any content in typical speech?
- Can naive listeners reliably detect the prosodic atypicality in HFA on the utterance level?
- How well can we predict perceptual ratings of prosody from acoustic-prosodic features?
- Is atypical prosody specific to ASD, or are certain atypical features also found in SLI?

1.3 Scientific Contributions

The contribution of this thesis include the following:

1. My work differs methodologically from most previous studies in that I analyze the prosody of spontaneous speech utterances while controlling for content features, such as the activity in which the child was engaged and the duration of the utterance. This allows me to examine the relationship between prosody and content.

2. I identified several acoustic-prosodic (objective) features that are significantly different in the speech of children with HFA from those with TD, or replicated such findings. First, several robust statistics of the fundamental frequency and intensity curves differ between the groups, especially higher F0 spread at the utterance level and higher variability of utterance-level statistical features. Second, several Functional Principal Component Coefficients, corresponding to curves that the intonation curve can be decomposed into, differ between the groups. Third, the intonation curve most characteristic of each speaker commonly starts from and continues at a higher pitch for subjects with HFA than for children with typical development.

3. I created a dataset of perceptual ratings of emotional and prosodic content for children's utterances.

4. The combined perceptual ratings for spontaneous speech utterances from multiple naive judges (on the order of five to ten raters per utterance) can reveal statistically significant differences between the groups with HFA and typical development. Namely, emotional arousal was generally higher in HFA than in TD, especially for utterances matched on content features. For a set of utterances with diverse content and prosody and not matched on content features, the raters perceived the speech of children with HFA as more atypical.

5. Atypical prosody, identified based on acoustic-prosodic features or perceptual ratings, is not peculiar to HFA with or without language impairment after controlling for content features, but is very similar in all aspects in SLI and is significantly different from TD, even for SLI subjects who do not display the features associated with autism.

Innovations in the area of computer science include the following:

6. I created an innovative approach and algorithms for matching multiple groups simultaneously on multiple groups of covariates.

7. I created a novel approach for estimating speaker-specific intonation curve characteristics making use of artificial training data synthesized with a text-to-speech system.

8. I created an iterative algorithm for estimating the bias and competence of raters for the situation when the number of ratings per rater was highly variable.

9. I was able to predict emotional arousal with a reliability comparable to human judges using robust statistical features of the F0 and intensity curves.

1.4 Organization of the dissertation

The next three chapters summarize topics required to understand our work. In Chapter 2, we overview some scientific concepts to put our research into perspective and to help the reader understand our contributions. We concentrate on the neurodevelopmental disorders under investigation here, especially Autism Spectrum Disorders (ASD), as well as speech prosody, and review previous findings from the literature on prosody in ASD. Chapter 3 explores the technical basis for analyzing the speech signal and modeling prosody, as well as the statistical repertoire that we use throughout the dissertation. Chapter 4 describes the speech corpora that we scrutinized to attain our findings.

The subsequent chapters are more or less independent pieces of our work, but later chapters often build on earlier ones. Chapter 5 provides one of the bases of our work: a novel methodology and tools for matching subject groups on multiple cognitive measures. Later chapters work with such matched groups. Chapter 6 compares the diagnostic groups on acoustic-prosodic features that capture statistical features of prosody as well as intonation curve shapes. In Chapter 7, we describe how we collected subjective ratings of prosody and emotions for utterances, assessed the data quality, and analyzed the data for group differences. In Chapter 8, we report results for predicting these perceptual ratings as well as those from another corpus using an acoustic-prosodic feature set. In Chapter 9, we summarize our results and draw conclusions about their significance. Finally, in Chapter 10, we outline our vision regarding where this line of research may lead over the course of multiple years. Let us start out by learning about the topic of this dissertation, the autistic condition.

Chapter 2

Background

2.1 Autism Spectrum Disorders (ASD)

Autism is an often serious neurological condition, characterized by persistent deficits in social communication and social interaction, and restricted, repetitive patterns of behavior, interests, or activities (American Psychiatric Association and others, 2013). The circumstances under which autism was first described seems to have had a lasting effect on the diagnostic practice and the research directions. Below we review the history of discovering this disorder to put our work into context. Then we relate its characterization and its relationship to other disorders.

2.1.1 History

The neurological condition termed "Autism" is generally considered to have been described first by two scientists independently in the first half of the 20th century: by Austrian–American child psychiatrist and physician Leo Kanner in 1943 (Kanner, 1943), who described 11 cases he had seen since 1938, and by Austrian pediatrician and medical professor Hans Asperger (Asperger, 1944b,a), who described four cases in his first paper on the topic of more than two hundred he had already seen.

The term "autistic" was introduced by influential Swiss psychiatrist Eugen Bleuler in 1910, and became widely known (so much so that Kanner, who himself had written a textbook on psychiatry, did not even feel it was necessary to explain it to his readers in his 1943 paper). It was used for describing one of the symptoms of schizophrenia, namely withdrawal from social relationships into their own world. This is the meaning in which Kanner and Asperger used the word "autism", essentially describing a condition similar in some of its features to schizophrenia, with the main distinction, as Kanner (1943) pointed out, that "in schizophrenic children or adults" there is "a departure from an initially present relationship", whereas in autism "there is from the start and extreme autistic aloneness." Asperger (1944a) similarly says that they "do not show the progressive deterioration that would be expected for psychosis. In essence, they remain the same throughout their life", and goes on to describe even improvements in social adaptation.

Both of them also emphasize the "monotonous repetitiousness and the resulting limitation in the variety of spontaneous activity" (Kanner), or "stereotypic activity", such as "movement stereotypies" and "monotonous play" (Asperger). Kanner and Asperger continue to describe a whole range of other behavioral markers of the children with these conditions, often the same phenomena in different words. The two we just pointed out, namely the social problems, together with restricted interests or repetitive behavior, have become the core diagnostic criteria for the condition.

The descriptions from Kanner and Asperger have many things in common: Both of them essentially say that there is something very peculiar about autism that is shared by all cases, while having outstandingly varying individual details. Moreover, both describe a spectrum of cases, from completely nonverbal and intellectually handicapped, to verbal and even smart children.

The two descriptions also have one major difference: Kanner states that "in none of the eight "speaking" children has language over a period of years served to convey meaning to others" and then they first just echoed sentences without any changes. In contrast, the cases described by Asperger use language at a high level from an early age for sophisticated topics, such as science and art, even though sometimes they use language in a peculiar way, using adult-like and newly invented words (neologisms). He does nevertheless mention that some of his cases are nonverbal. Today we believe that both scientists described essentially the same phenomenon, only Kanner's cases were all low-functioning, while Asperger concentrates on high-functioning children and makes only one mention of lower-functioning individuals. (Perhaps he did this out of humanitarian motives, as for example Uta Frith, who published an English translation of Asperger's paper, suggests (Asperger, 1944a): "The historical background to this passionate defence of the social value of autism was the very real threat of Nazi terror which extended to killing mentally handicapped and socially deviant people.")

Nevertheless, these two authors considered the disorders they identified as different: Kanner states in a book review in 1970: "Asperger ... independently described what he called *autistic psychopathy* which, if at all related to infantile autism, is at best a 42nd cousin and merits, and has received, serious attention from investigators not confused by klang association." (Kanner, 1970) Asperger also maintains, as Uta Frith summarizes, that "these cases are very different in that there is a different aetiology. Severe cases of autism, in his view, had to have brain insult rather than constitutional causes." (Asperger, 1944a) The two conditions were indeed treated as related but separate entities for a long time, and were merged into a common diagnostic category starting with DSM-5, the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association and others, 2000).

It may be worthy of note that the fact that Kanner's description was used as a basis for diagnosis in the US seems to have had a lasting effect on the landscape of autism. Even in 1966, a large research project on autism (Pronovost et al., 1966), which studied 13 autistic children, only had low-functioning cases, so much so that even the verbal subjects "were seriously impaired in their capacity to comprehend language." With today's diagnostic criteria, many of those considered autistic are clearly high-functioning, so some of the increase in the number of cases is obviously due to the changed diagnostic practice (Whitehouse et al., 2017).

Russian child psychiatrist Grunya Ssucharewa (Ssucharewa, 1926) described six boys, aged 2 to 14 years in 1925, whom she considered to have the "schizoid personality disorder" described by Kretschmer (1922). Wolff (1996) published its English translation and claims that Ssucharewa may have described cases with Asperger's syndrome. Note that she did not think she had discovered a new condition, nevertheless the cases she describes seem to be very similar to the ones described by Asperger almost two decades later. For example, she says her subjects have an "autistic attitude". She occasionally also refers to the other defining characteristic of autism, namely restricted interests and repetitive behaviors. She mentions its presence for a few her cases, and seems to give importance to it when she contrasts the syndrome with typical development: "Isolated schizoid features are not infrequently seen in normal children who often grimace, repeat words stereo-typically, and invent new words."

Lately, it has even been questioned if autism should remain a diagnostic entity (see e.g. Waterhouse, 2016), or if instead it should be split into two or three entities (related to restricted and repetitive behaviors, and problems with socialization and communication), or should be defined based on biomarkers (see e.g. Waterhouse and Gillberg, 2014). While it is accepted that ASD has a wide range of etiologies and neurological backgrounds (Happé et al., 2006), it is also to be expected that the diagnostic category of autism is to stay with us, at least for a very long time (Happé, 2017).

2.1.2 Characterization

Autism is diagnosed based on one of two sets of criteria: The DSM, used mainly in the US, and the International Classification of Diseases (ICD), used mainly in Europe. The latest version of the latter is ICD-10, which is very similar to DSM-IV-TR (American Psychiatric Association and others, 2000). For our purposes, DSM-IV and DSM-5 (American Psychiatric Association and others, 2013) are the most important ones: DSM-IV because the subjects whose data we have been working with were diagnosed based on its criteria, and DSM-5 as the lastest version. These two have a very similar definition of the condition. The exact diagnostic criteria can be found in other publications, here we just summarize the main characteristics:

- DSM-IV
  - qualitative impairments in social interaction
  - qualitative impairments in communication
  - restricted repetitive and stereotyped patterns of behavior, interests and activities

  - onset prior to age 3 years

- DSM-5

  - persistent deficits in social communication and social interaction
  - restricted, repetitive patterns of behavior, interests, or activities
  - moreover, the symptoms together limit and impair everyday functioning and must be present in early childhood

Research on autism can be valuable because the prevalence of autism is relatively high, with the most likely estimate being 1 in 152 children (Hill et al., 2015b), although estimates vary widely, partly due to changing diagnostic criteria, increasing awareness, and unequal access to healthcare. It is much more often diagnosed in boys than in girls. It is an idiopathic disorder, in other words, its causes usually cannot be identified with certainty. Some genetic, epigenetic, and environmental risk factors have already been identified (see e.g. Hallmayer et al., 2011).

It has great impact on quality of life, both for the individual and his or her family (Allik et al., 2006; Lee et al., 2008). This has financial consequences as well: It is very expensive both to the family and the whole society. According to one estimate, autism costs a staggering $3.2 million per capita in the USA, mainly due to lost productivity and adult care (Ganz, 2007).

On the positive side, early intensive behavioral intervention by therapists and even by parents is often effective in improving debilitating symptoms of autism (see e.g. Reichow et al., 2011; Green et al., 2015)). This makes it all the more important that it is possible to diagnose autism by the age of two (Lord et al., 2006). Unfortunately, it is still often diagnosed much later. Early intervention requires that all children get appropriate early screening, as even professionals miss over a third of the autism cases when just observing children for a limited amount of time (Gabrielsen et al., 2015). Current diagnostic procedures are time-consuming and require the presence of trained clinicians, hindering the introduction of population-wide screening. Computational aids to prosody assessment may alleviate the problem by providing automatic or semi-automatic means to early screening from infant vocalizations, and there are already some promising approaches (see e.g. Santos et al., 2013).

2.1.3 Relationship to other disorders

The DSM-5 definition of autism subsumes syndromes formerly classified as separate disorders, such as Rett's Syndrome and Asperger's Syndrome. ASD is often comorbid (co-occurring) with other conditions, especially Language Impairment (LI), Intellectual Disability (ID, formerly called

Mental Retardation), and Attention-Deficit Hyperactivity Disorder (ADHD). Around 50% of all children with ASD also have language impairment (Loucas et al., 2008; Leyfer et al., 2008). As many as 40% have ID, of which over 20% is borderline ID. About 30% is without speech at age 4, although most of them learn to speak eventually (Wodka et al., 2013).

2.2 Speech Prosody

The word "prosody" originally referred to a song sung to music, and later came to mean the music of human speech as well. It is what is sometimes called the "tone of voice" in everyday speech. It is expressed through the melody, the rhythm, pausing, and loudness variation of speech, as well as the timbre or voice quality, such as hoarseness or nasality. These are expressed acoustically using the fundamental frequency (F0), segmental (phoneme) durations, speech intensity, and glottal and spectral characteristics of the voice. The importance of voice quality for expressing meaning varies from language to language; it is vital for example for the Japanese language (see e.g. Campbell and Mokhtari, 2003; Ito, 2005). In phonetics, it is also called the suprasegmental level of speech, which is superimposed onto the segmental level, that is, the words and speech sounds of each sentence. We distinguish it from intonation, which we only use for the variations of the fundamental frequency, that is, the melody.

Speech prosody carries various kinds of information, and accordingly it contributes meaning in various ways to the listener. We control prosody to express our thoughts better, and our actual physical and mental state also affects it involuntarily. Due to its various functions, prosody is important for child language acquisition (Sharifi et al., 2016). We briefly describe these functions and roles, not primarily to educate the reader but to put our work into context; for a more detailed description, see e.g. Peppé (2009). Note that the boundaries of some of these categories is not carved into stone: One could organize these aspects of prosody somewhat differently.

- Lexical function: In tonal languages, so-called "lexical tones" affect the meaning of the phoneme sequences. In the English language, the lexical accent similarly distingishes meaning for certain words, such as REcord (stored information) vs. reCORD (the act of storing the information).

- Grammatical function: Prosody can serve to help disambiguate syntactic elements of speech, such as sentence type (question vs. statement) or the presence of phrase boundaries.

- Pragmatic function: It includes indicating the focus of the utterance, emphasizing given vs. already known information, and the attitude and intent of the speaker, such as whether

s/he wants to keep or yield the floor to the conversation partner (turn-taking), moreover "negotiating agreement, signaling recognition and comprehension, and managing interpersonal relations such as control and affiliation" (Ward, 2004).

- Affective role: Prosody carries information about one's emotional states and mood, to some extent even when one would rather hide it. One can also pretend certain emotional states to varying degrees, for example to express sarcasm.
- Indexical role: It is not just voice timbre and pronunciation that is characteristic of the speaker, but also one's individual prosody. With enough experience, one can even recognize the group the speaker belongs to, such as dialect, geographical area, cultural background, or even a particular family, just based on prosody if it is pecullariar enough.

Prosodic form and function are related in a complex way by necessity, as all these functions need to be expressed by variations of the aforementioned acoustic features: F0, intensity, and segmental durations, and to some extent spectral aspects of speech, such as voice quality. Both the statistical properties of these acoustic features and their changes in the course of time can be important. It may be possible to express a function of prosody using multiple forms, and vice versa, a prosodic form may correspond to multiple functions.

The indexical role is not independent of the other roles of prosody: It seems possible that when a certain prosodic function can be expressed in multiple ways, the proportion of each particular way of realization is characteristic of the speaker's group membership. It may also be the case that non-functional prosodic elements are superimposed on the functional part, both identifying the speaker's group and giving a particular (occasionally unnatural) flavor to his or her speech.

2.2.1 Prosodic disorders

Scientists have delineated multiple Spoken Language Disorders (SLD; American Speech-Language-Hearing Association, 2016). When SLD is not comorbid with other disorders (such as hearing impairment or another sensory disorder, global developmental delay, intellectual disability, etc.), it is called *Specific Language Impairment* (SLI; Tomblin, 2011); other names include Developmental Language Disorder, Language Delay, and Developmental Dysphasia (National Institute on Deafness and Other Communication Disorders, 2016). SLI is significant for this study as it forms one of our control groups. SLD is sometimes (or maybe always, it is not yet clear) accompanied by deficiencies of speech prosody as well.

Clinicians have found peculiarity of speech prosody in various disorders, with or without other language problems. Zei Pollerman (2002) has posited that prosody is an essential mechanism

for investigating cognition and emotion. Dysprosody is a specific deficit in prosody affecting one or more of its functions. Aprosodia is dysprosody in the area of emotional communication only, that is, a deficit in expressing or interpreting emotional prosody. But atypical prosody has been noted in other disorders as well where it is not the defining characteristic, only accompanies the fundamental deficit, at least for some of the subjects (Cleland et al., 2010). We cannot preclude the possibility that prosody is affected in a disorder-specific way.

Some of the disorders accompanied by atypical prosody or possibly dysprosody are:

- Autism Spectrum Disorders (ASD) (Shriberg et al., 2001),
- Specific Language Impairment (SLI) (Stojanovik and Setter, 2009; Demouy et al., 2011),
- Intellectual Disabilities (Shriberg and Widder, 1990; McSweeny and Shriberg, 2001),
- Apraxia (Odell et al., 1991; McSweeny and Shriberg, 2001),
- Hearing impairment (Parkhurst and Levitt, 1978),
- Down's syndrome (Stojanovik, 2011),
- Williams Syndrome (Stojanovik and Setter, 2009),
- Alzheimer's disease (Roberts et al., 1996),
- Parkinson's disease (Darkins et al., 1988),
- Schizophrenia (Murphy and Cutting, 1990),
- Dysarthria (Odell et al., 1991; Bunton et al., 2000),
- Aphasia (Moen, 2009),
- Traumatic Brain Injury (Angeleri et al., 2008),
- Depression (Alpert et al., 2001).

In this work, we concentrate on ASD, with SLI and TD as control groups.

2.2.2 Screening instruments for prosody

There are two screening instruments that have been developed for assessing prosody for children: One is the *Prosody-Voice Screening Profile* (PVSP) created by Shriberg et al. (1992), the other is *Profiling Elements of Prosodic Systems–Children* (PEPS-C) created by Peppé and McCann (2003)

and adapted for children from a generic version (Peppé et al., 2000). We briefly summarize these below for easy reference. For an in-depth discussion of screening tools, see e.g. Diehl and Paul (2009). Let us point out here just one of their remarks that is very important to our work: Unlike for language assessment tools, currently there are no screening instruments that have been standardized on a large subject pool. So we cannot know for example what amount of atypical samples are acceptable in a child's speech such that overall it can still be considered typical.

2.2.3 Prosody-Voice Screening Profile (PVSP)

The PVSP screening instrument, designed by Shriberg et al. (1992), characterizes prosody through perceptual ratings of phrasing, speech rate, stress, and three additional factors they call "voice characteristics", namely loudness, pitch, and voice quality (Shriberg et al., 1992, p.45, Table 1). The raters can choose the "appropriate" label, or a characterization of how the speech deviates from the expected standard, such as "too slow" for speech rate. Paul et al. (2005b) describes PVSP the following way (quoting verbatim from page 864):

- "Phrasing: the smoothness or fluency of speech (part- and whole-word repetitions and revisions)."
- "Rate: the overall pace of speech (too slow or too fast, as measured by syllables/second)."
- "Stress: the relative emphasis on syllables and words (intensity, pitch, duration)."
- "Loudness: the intensity with which utterances are produced (too loud, too soft)."
- "Pitch: the average frequency of the voice (too high, too low)."
- "Voice quality: the sound produced by the larynx (e.g., harsh, strained)."
- "Resonance: the sound produced by the vocal tract (e.g., nasality, denasality, pharyngeal resonance)."

The ratings are applied to spontaneous speech utterances, after excluding certain kinds of utterances, such as short utterances (defined as less than four words), back channels, imitations, repetitions, names, reading, singing, too many unintelligible words, whisper, sound effects, and noise (environmental, line noise, and body sounds, such as laughing or sneezing; see Shriberg et al., 1992, Figure 1, Exclusion Codes). The judges need to learn from the PVSP training material, which takes an average of 15 hours to acquire (ranging from 10 to 18 hours). Even with this amount of training, code-level reliability (inter-rater agreement) may not be adequate for certain measures in the PVSP profile, especially when a code occurs relatively infrequently (Paul et al., 2005b).

2.2.4 Profiling Elements of Prosodic Systems–Children (PEPS-C)

PEPS-C is an application for examining both expressive and receptive abilities of prosody in regard to certain functional aspects. It is quite different from PVSP in that it is for assessing the prosodic abilities of children through computerized tasks, whereas PVSP does that by annotating speech utterances.

The functional aspects assessed by PEPS-C are:

- turn-end type (question vs. statement),
- affect (liking vs. disliking),
- chunking (phrase boundaries),
- contrastive stress,
- lexical stress (only in the 2015 version), and
- phrase stress (only in the 2015 version).

It also includes form-tasks:

- auditory discrimination task, and
- imitation task.

For receptive tasks, the subject selects one of two pictures that correspond either to the correct answer or to a distractor, and the task interface records his or her score. For expressive tasks, the examiner rates the child. For example, the imitation tasks are evaluated on a three-point scale (good, fair, poor). The test does not involve spontaneous speech: Not just the receptive tasks have fixed contents, but the expressive tasks also require the children to say certain pre-determined expressions. The test is available for several dialects of English and some other languages.

2.3 Speech Prosody in ASD

From the earliest descriptions of autism, clinicians noted that the speech prosody of autistic people often sounds atypical or strange. These differences are beyond the general variation expected based on cultural, gender, and personal differences. Prosodic deficits in autism are closely related to one of the two core symptoms of autism: impaired social communication. Atypical expressive prosody can put a stigma on people living with autism, besides carrying the risk of making them misunderstood,

while impaired receptive prosody will of course make it harder for them to understand others. Their prosody can sound inappropriate, even rude (McCann et al., 2007; Peppé, 2009), and prosodic deficits are correlated with social and communicative abilities (Paul et al., 2005b). This implies that these problems can hinder their social acceptance.

Nevertheless, prosody is not included in diagnostic procedures. Even though prosody is scored in the ADOS (see Section 4.2.1), this score is not taken into account in the diagnostic algorithm. Prosodic disorder is not part of the diagnostic criteria for autism, probably because it is possible that deficits are not present for certain individuals, and prosodic disorder is also not specific to autism as we have seen earlier (see Section 2.2.1). Yet examining it could help to confirm the diagnosis in certain cases. It is probably not included due to a lack of prosodic tests that have ecological validity, are based on empirical information, have a representative normative sample, and are easy to administer (Diehl and Paul, 2009).

In the following, we look at what researchers have already found out about prosody in autism: qualitative differences in its subjective perception and its objective characterization, as well as functional deficiencies.

2.3.1 Early mentions

The seminal papers on autism by Kanner (1943) and Asperger (1944b), as well as the one by Ssucharewa (1926), already mentioned atypical prosody as a characteristic of the speech of individuals with autism. They include this feature in their summaries about autistic behavior, yet it is not obvious that they considered it as a defining characteristic, as Kanner and Ssucharewa do not mention atypical prosody as a symptom for each of their subjects. Since these papers have been cited a great deal in connection to the suprasegmental aspects of speech as well, namely prosody and voice quality, let us look at them in detail. As we shall see, they have already touched upon all the phenomena qualitatively that later papers described in more detail. (The emphases in the text are from me.)

Kanner

Kanner mentioned atypical prosodic aspects for 5 of his 8 speaking subjects (Kanner, 1943):

- Case 1: "He used the personal pronouns for the persons he was quoting, even **imitating the intonation**."
- Case 4: "...these utterances, made **always with the same inflection**..." (It is obvious from the context that by the word "inflection", he refers to intonation.)

- Case 7: "He sometimes uttered inarticulate sounds in a **monotonous singsong manner**."
- Case 8: "He **never asks questions in the form of questions** (with the appropriate inflection)."
- Case 11: "She speaks well on almost any subject, though with something of an **odd intonation**."; "She reads very well, but she reads fast, jumbling words, not pronouncing clearly, and **not making proper emphases**."

In his discussion, he notes some common characteristics: Regarding echolalia that "Not only the words, but even the **intonation is retained**." When children are left alone, there is sometimes "happy though **monotonous humming and singing**." And "the children **read monotonously**." In summary, he pointed oddities of intonation, emphases, and a monotonous way of speaking, and mentioned such details for 3 of his 8 speaking subjects.

Several later papers pointed out that autistic speech has been associated with both "monotonous" and "singsong", saying that there is an apparent contradiction. Note though that in Kanner's paper these words appear together: "monotonous singsong manner". This may refer to what some studies have found: A relatively small set of singsong intonation patterns repeated again and again.

Asperger

Asperger described four patients he chose from around two hundred (Asperger, 1944a), going over essentially the same topics for everyone (family history, appearance and behavior, intelligence testing, behavior on the ward, etc.). He mentions somewhat peculiar speech and irregular prosody in each case:

- Case 1 (Fritz V.): "...his voice, which was **high and thin** and sounded far away. The normal speech melody, the natural flow of speech, was missing. Most of the time, he spoke **very slowly**, dragging out certain words for an exceptionally long time. He also showed **increased modulation** so that his speech was often **sing-song**."
- Case 2 (Harro L.): "His voice ... was **very deep** and appeared to come from very far down, the abdomen. He talked **slowly** and in a deadpan way **without much modulation**."
- Case 3 (Ernst K.): "His voice ... high, **slightly nasal** and **drawn out** ..."
- Case 4 (Hellmuth L.): "...talked **slowly**, almost as if in verse"

In his summary of the clinical features, he says: "Sometimes the voice is **soft and far away**, sometimes it sounds **refined and nasal**, but sometimes it is **too shrill and ear-splitting**. In yet

other cases, the voice drones on in a **sing-song** and does not even go down at the end of a sentence. Sometimes speech is **over-modulated** and sounds like exaggerated verse-speaking. However many possibilities there are, they all have one thing in common: the language feels **unnatural**, often like a caricature, which provokes ridicule in the naive listener." In summary, he found atypical voice quality, usually too varied intonation or in one case almost no intonation, inadequate control of the intensity, and slow speech. He seems to indicate that all of his cases had some unnaturalness in their prosody.

Ssucharewa

Ssucharewa described six cases of "schizoid personality disorder" (Wolff, 1996), whose properties resemble Asperger's syndrome.

- Case 1: "speech **lacking in modulation**"; "speech **rapid** but unclear."
- Case 2: "**monotonous** voice"; "**without any change of intonation**"; "voice **nasal**", and again: "**nasal** speech."
- Case 4: "**often pauses** in his flow of speech"; "**adult intonation**"; "**high pitched, whiney** voice."
- Case 5: "**deep, hoarse** voice."

Later she sums up these as "**oddities and lack of modulation** of speech" and "peculiarities of voice and language". In summary, she mentioned atypical voice quality (in two cases), monotonous or sing-song type of intonation (three cases), and too rapid speech and unexpected pauses once. So she observed such phenomena frequently, if not universally.

2.3.2 Overview of papers on prosody in autism

Let us now move on to what later studies have found on prosody in autism, following these early mentions and other clinical observations. Before going deeper, first we will look at some summary statistics on the studies.

In Figure 2.1 we can see how the total number of papers that deal substantially with prosody in autism accumulated over the years (this includes papers that just summarize existing findings; the cummulative count plots were derived from our paper summary table). Until around 1980, such papers were few and far between, then their number increased at a higher rate. In 2003, McCann and Peppé (2003) published a critical review on research on prosody in autism, which

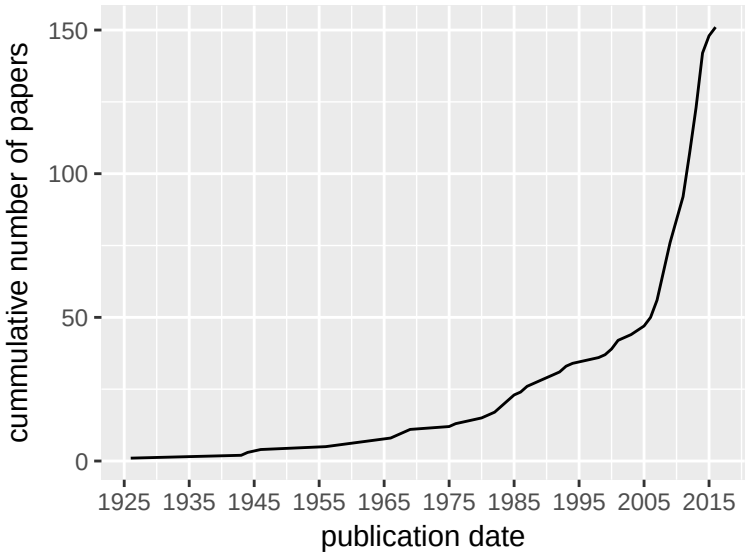


Figure 2.1: Cummulative number of papers by year that substantially deal with prosody in autism

became a highly cited paper. They concluded that "prosody in autism spectrum disorders is an under-researched area and that where research has been undertaken, findings often conflict." Perhaps it is not a coincidence that following its publication, there was a steep rise in the number of publications on the topic, starting with 2005. McCann and Peppé (2003) also said that "Only two studies (Baltaxe et al. 1984, Fosnot and Jun 1999) use acoustic analysis to quantify expressive prosody (seven do not); more is needed to establish the prosodic features that characterize both atypical and typical prosody." As we can see in Figure 2.2, starting with 2007, the number of such studies started to increase substantially. Some of that increase came from Peppé and her colleagues, who utilized the PEPS-C test of prosody for autism.

2.3.3 Perceptual ratings

A number of studies have collected perceptual ratings of prosody using the PVSP or the PEPS-C tasks, or some other one-off task. These are somewhere between qualitative and quantitative characterization, as they reflect subjective judgments, but in a way that makes it possible to characterize those quantitatively based on an (often large) array of subject $\times$ task $\times$ rater scores. We summarize the results of some of these studies below, separately for each rating task.

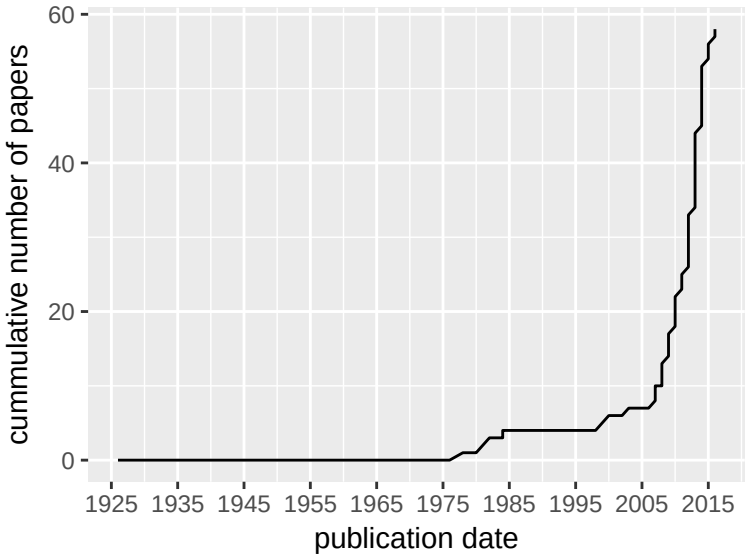


Figure 2.2: Cummulative number of papers by year that do acoustic analysis of prosody in autism

2.3.4 PVSP studies

Shriberg et al. (2001) used the PVSP (see Section 2.2.3) to analyze the speech of 15 subjects with HFA, 15 with Asperger's Syndrome (AS), and 53 TD subjects, aged 10 to 49, a relatively wide age-range. One transcriber rated all the recordings and 20% was rated by another one to calculate interjudge agreement, which was 93% for exact agreement. They said that a subject had failure scores for a prosodic variable if more than 20% of his or her utterances were rated as atypical. They found that significantly more HFA and AS speakers failed on stress and voice quality, and significantly more AS speakers failed on phrasing.

In her MSc thesis, McAlpine (2012) (see also McAlpine et al., 2014) analyzed the PVSP scores for 7 subjects with ASD and 7 TD controls, aged 2 to 6. Two or three investigators scored the utterances, with an average of 94% interjudge agreement. The author found significant effects of prosody type and group on the percentage of inappropriate ratings, with no significant interaction. Overall, the ASD group tended to have more atypical ratings, mostly due to atypical stress patterns, such as misplaced or reduced stress.

In summary, the findings found that the ASD group sounded more atypical than the TD group, with one consistent finding, namely that stress is compromised in ASD. Other atypical aspects that were sometimes indicated are voice quality and phrasing.

2.3.5 PEPS-C studies

McCann et al. (2005) asked children with ASD and TD matched on Verbal Mental Age (VMA) to complete the PEPS-C (see Section 2.2.4). They found that the autism group performed significantly worse. All children in the ASD group had problems with at least on of the tasks. Unfortunately we do not know further details because only a summary of this paper is available.

Peppé et al. (2006, 2007) and McCann et al. (2007, 2008) worked with 31 high-functioning HFA and 72 TD subjects. The age range was 6 to 13 years for the HFA group and 4 to 11 for the TD group. The groups were matched on VMA; as a result, chronological age (CA) was significantly larger in the ASD group. The groups did not differ significantly on gender and postal code, a proxy for socio-economic status. The HFA group performed significantly worse on 7 of the 12 tasks: the affect-tasks, item discrimination and imitation, and focus.

McCann et al. (2008) inspected the effect of language abilities on prosodic abilities on a group of children with ASD with or without a comorbid language impairment, assessing the functional aspects of prosody using PEPS-C. She found that prosodic abilities correlated with language abilities, yet autism itself also had an effect on these abilities, refuting the supposition that prosodic deficiencies can be accounted for purely as the consequence of issues with language.

Peppé (2006) reported the results of the PEPS-C for 29 children with HFA, 29 with AS (with no clinically significant pre-school language delay), and 25 TD children aged 5 to 14, matched on VMA, with significantly larger CA in both ASD groups. The HFA group scored significantly lower than both of the other two groups on affect, focus, imitation of words, and imitation of phrases, and additionally lower than the TD group on turn-end and chunking. The AS group scored significantly lower than the TD group only on imitation of phrases. They also acquired atypicality ratings for 10 seconds of conversational speech, both from 5 "phonetically aware judges" and from 22 "naive judges". They did not specify how they selected the conversation speech samples. The atypicality ratings had a moderate correlation (0.57 and 0.61) with the PEPS-C scores.

Peppé et al. (2011) compared the prosodic performance of children with HFA (31) or AS (40), and both CA-matched and Lexical Mental Age-matched (LMA) TD children where LMA is estimated based on the vocabulary size using the British Picture Vocabulary Scales (BPVS-II; Dunn and Dunn, 2009). They found that HFA performed worse than CA-matched controls on all measures (affect, sentence type, contrastive stress, phrasing, imitation), but not significantly worse on two measures (sentence-type, phrasing) compared to LMA-matched controls. AS performed worse in both cases on phrasing and imitation only. They found significant differences on the expressive language skills as well, but they conjectured that the prosodic differences cannot be

accounted for by these differences.

Järvinen-Pasley et al. (2008) related the performance of 21 children with ASD (AS or Kanner's autism) and 21 with TD to each other on form- and function-level tasks, using both the PEPS-C and a sentence type discrimination task, where the children were to decide whether a sentence with manipulated prosody was a question or a statement. They found that all children performed adequately "in both form and function tasks at the single-word level", but the children with ASD had significantly worse performance on the sentence-level tasks. Moreover, those with ASD showed a bias toward misrecognizing questions as statements.

Diehl and Paul (2011, 2012) elicited speech from children with ASD (24), Learning Disabilities (LD, 16), or TD (22) using the PEPS-C tasks. The children were between the ages of 8 and 17. They found that both the ASD and the LD groups performed significantly "worse at perceiving and imitating prosodic patterns than the TD comparison group."

Filipe et al. (2014) had 29 subjects speaking European Portugese, namely 12 AS and 17 TD children aged 8–9, perform the turn-end task of PEPS-C. They did not find significant differences on this subtask. Filipe (2014) gives acount of another experiment wherein she asked 15 HFA subjects aged 5–9 and 15 TD children matched on age and non-verbal IQ (NVIQ) perform all PEPS-C tasks. The HFA children generally performed worse on the tasks, with statistically significant differences for imitation, affect expression, turn-end expression and comprehension, and auditory discrimination. We review their other findings in Section 2.3.6.

Hesling et al. (2010) administered the French version of PEPS-C and PEPS (the adult version) to HFA and TD subjects and compared its results with that of their Functional Magnetic Resonance Imaging (fMRI) data. They found that the brain system that is respondible for perception of speech prosody is activated atypically compared to the controls.

2.3.6 Other perceptual studies

A number of studies came up with (or re-analyzed data for) a new evaluation task to shed light on one specific issue and asked multiple judges to answer those questions, often asking for a rating on a (continuous or discrete) scale (Shriberg et al., 2001; Paul et al., 2005b; Peppé, 2006, 2007; Hubbard and Trauner, 2007; Le Normand et al., 2008; Paul et al., 2008; van Santen et al., 2010; Diehl and Paul, 2011; McAlpine, 2012; Nadig and Shaw, 2012; Grossman et al., 2013; Bone et al., 2015; Kaland et al., 2013; Filipe et al., 2014; Filipe, 2014; Sharifi et al., 2016). Below we review some of these.

Peppé (2006) acquired two sets of perceptual atypicality ratings using Direct Magnitude Estimation (Campbell and Dollaghan, 1992) for 10 seconds of conversational speech for the speech

of children with HFA, AS, or TD (see more details in Section 2.3.5). The atypicality ratings were significantly different for all three group-pairs except that only the scores of the phonetically aware judges were significantly different for the HFA–AS comparison, not those of the naive judges.

van Santen et al. (2009) adapted PEPS-C and two tasks created by Paul et al. (2005a). The latter comprised a lexical stress task, a pragmatic style task, and an affect-task, wherein the child was to say the phrase "It doesn't matter" imitating one of four emotions at a time (happiness, sadness, anger, fear). Later they utilized three of these tasks (van Santen et al., 2010) for a group of children with HFA (26) or TD (26) matched on age and NVIQ. Two of the three automated tasks were able to differentiate between the ASD and TD groups, that is, the scores calculated from tasks were significantly different between the groups. They also found that it was not the ability of the participants to express functional differences that differentiated the groups, but the balance in the use of prosodic features, namely F0 and duration: The children in the ASD group tended to rely mostly on F0 to express contrasts, whereas the TD group tended to use both F0 and duration. This is similar to what Baltaxe (1981) found regarding disproportionate use of different modalities. Nadig and Shaw (2015) also described a non-functional difference in the use of prosodic features: They compared the marking of prominence in the speech of preadolescents with HFA or TD, and found that neither group relied on pitch, with "HFA speakers relying less consistently on amplitude" and TD "speakers relying less consistently on duration."

Nadig and Shaw (2012) recorded conversation with children with HFA (15) or TD (13) aged 8 to 14. They played 10-13 seconds from the longest uninterrupted speech segment of each child (generally two to three utterances) to Speech-Language Pathology students, who rated the segments on four rating scales:

- pitch height on a 7-point scale, from "low" through "normal" to "high";
- pitch changes on a 7-point scale, from "flat / monotone" through "normal" to "too variable / sing-song"; the raters were previously shown reference examples of flat vs. too variable pitch;
- speech rate on a 7-point scale, from "slow" through "normal" to "fast";
- overall typicality on a 4-point scale from "atypical" (1) to "normal" (4).

They found that 9 of the 15 children with HFA had typicality scores below the range of the TD group (or to put it another way, 6 children with HFA were in the range of the typical children and 9 were below that). The mean typicality score of the HFA group was significantly lower than that of the TD group (2.76 vs. 3.23 on a 1 to 4 scale). There were no other significant differences.

Filipe et al. (2014) (see also Filipe, 2014) asked 35 adult listeners whether the utterances produced by children with AS sounded natural or odd. See a description of the subjects above in Section 2.3.5. The raters listened to all 16 one-word utterances for each child, followed by a 10 second pause before the recordings for the next child for rating what they had heard. They were asked to judge the atypicality of the utterances on a five-point scale, from common (1) to uncommon (5). They were blind to the purpose of the study. The children with AS had significantly higher atypicality scores.

2.3.7 Relationship to intellectual disability

As noted earlier, ASD co-occures with ID in a substantial number of cases (see e.g. Matson and Shoemaker, 2009), in up to 70% of the cases according to some estimates. Mayes and Calhoun (2011) conducted an analysis of 777 autistic children to determine the relationship of autism severity and autism symptoms to IQ and some other measures. The children came with a wide range of ages (1 to 17 years) and cognitive abilities (IQ between 9 and 146). The symptoms were assessed using the *Checklist for Autism Spectrum Disorder* (CASD; Mayes et al., 2009), taking into account not just parent interviews but other available records as well when determining the scores. This checklist has only one item on speech atypicality, namely "atypical, repetitive vocalization or speech". This item can obviously refer to a range of vocal phenomena, including echolalia and atypical prosody. They found that the prevalence of this symptom — which was rather high, in the range of 86% to 92% — was not correlated with IQ. This suggests that both high-functioning and low-functioning individuals show such atypicalities, moreover that these issues do not constitute a developmental delay, but rather persist over time.

2.3.8 Developmental aspects

Snow and Balog (2002) give a detailed overview of the prosodic development of typical children from infancy to about two years of age. Main points include the following. Even babies already use precursors of intonation from around the age of three months; starting around nine months, form–meaning associations disappear through regression, later to re-appear as a consequence of experience with their native language with single word utterances. The first controlled components of intonation are *register* (the overall pitch height for an utterance), then contour direction. The use of falling and rising contours to mark utterance boundaries seems to develop without regression starting with babbling; falling intonation is well developed by the age of one, whereas the use of rising intonation continues to develop through the preschool years.

Ballard et al. (2012) discuss one aspect of typical prosodic development between the ages of three and seven, namely lexical stress in multisyllabic words, which is deficient in several disorders. They found that (at least in Australian English) this contrast is expressed through the use of intensity and vowel durations, and even seven-year-olds cannot produce the distinction with the accuracy of an adult, probably due to physiological change. Literature they cite states that oral-motor production continues to develop beyond age 14.

Since prosodic development takes so long even for neurotypical children, it is no wonder if it takes even longer for children with neurodevelopmental disorders, occasionally with persistent deficits in certain areas. Sheinkopf et al. (2000) found that infants with autism do not have problems with producing well-formed syllables, but they produce significantly more syllables with atypical voice quality, an aspect of prosody. Shriberg et al. (2011) suggest that in autism, the framework for fine-tuning speech qualities including prosody may be "selectively impaired while the ability to tune in to the acoustic features may not be."

2.3.9 What is still not known

Studies on prosody in autism approached the topic from multiple perspectives, but some questions remain unanswered or are rarely studied.

We know very little about how prosodic ability and possibly atypicality of prosodic expression in SLI relates to that observed in ASD. Stojanovik and Setter (2009) summarized the literature on prosody in SLI saying that the evidence is not conclusive, but prosodic deficits seem to be present for at least some individuals with SLI. So far, however, we know of just one study where prosody in ASD and SLI was compared, and that only on an imitation task: Demouy et al. (2011) asked children to imitate words and sentences with descending, falling, floating, or rising intonation, and automatically evaluated their success. They found that children with SLI were very similar to those with ASD or Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS) and performed the sentence tasks worse than those with TD, except for rising intonation which they recognized and imitated similarly well to those with TD. Other than that, we know very little about this relationship.

There are very few studies that assess atypicality of prosody in detail at the utterance level. The numerous PEPS-C studies approach prosody through its functional aspect, whereas there have been only two PVSP studies, which characterize its atypicality. Other than that, most researchers use longer speech segments or multiple utterances to get one atypicality rating, and as a rule, find a difference compared to typical development. It is obviously much easier for listeners to judge prosody based on longer segments, simply due to having a longer speech duration (see e.g. Schuller

and Devillers, 2010), but it does not tell us what the reason for the perceived atypicality is. One data set goes deeper than that for read speech (Grossman et al., 2013; Bone et al., 2015), but its granularity is far from that of the PVSP. None of these examines spontaneous speech utterances in detail.

We do not know of any research aiming to characterize prosody in spontaneous speech that takes content features into account as well. The activity the child is engaged in, the topic, the sentence structure, the meaning all play a role in forming prosodic expression, but we do not know this relationship in detail. We also do not know of any studies that looked at how the intonation curves are different qualitatively in autism.

Beyond all that we have looked at, replication of the findings by other researchers and on other corpora is in itself important. There is no substitute for this in determining if a result is reliable, as of course significant findings can be due to chance. Beyond theoretical considerations, the ultimate test of whether a finding is externally valid is also replication.

2.3.10 Clinical practice

Crystal (2009) summarized the status of prosody in the introduction in a special issue of the International Journal of Speech-Language Pathology: Prosody was considered "the Cinderella of the linguistic sciences" for a long time, and "when it comes to the application of these analyses to the understanding of prosodic disability, [prosody] remains in the cellar, with few visitors." He states that prosody is often disregarded in diagnosis, assessment, and treatment, and few people try to overcome the practical difficulties to incorporate analyses of prosody. We hope that the accumulating knowledge on the usefulness of subjective and objective measures of prosody will find their way into clinical practice in the form of standardized tools with normative data.

Chapter 3

Technical Background

Now that we have learned about the scientific background, we review some computational techniques in this chapter that we employed in this work. The techniques described are generally not our own contributions (with the exception of some elements in the F0 tracking section), just a summary and synthesis of existing knowledge. Our purpose in this is to make this dissertation more accessible to those so far unfamiliar with these techniques, and to establish the meaning of how we use certain concepts.

3.1 F0 tracking

Perhaps the most salient (and most often analyzed) feature of prosody is pitch and its acoustic correlate, the fundamental frequency (F0). To be able to analyze F0, first we need to determine a time-series of voicing decisions (voiced or unvoiced) and F0 values for voiced regions of the speech segment being analyzed using an *F0 tracker* (also called *pitch tracker*). These algorithms generally determine the F0 curve using a sliding window for overlapping speech frames of a predetermined length (called frame size or window size) that are taken once every N milliseconds (called frame shift or window shift). The frame shift parameter defines the level of detail captured. The frame size parameter sets a lower bound on the F0 range that can be determined; this lower bound can be decreased only at the expense of losing fine details of the curve. Setting these parameters and the expected range of F0 values correctly for the particular speaker is essential for detecting the correct curve (see e.g. Evanini et al., 2010). But even if these parameters are optimal, currently no F0 tracker produces an error-free output for all voice types and conditions (Kawahara et al., 2005; Ewender and Pfister, 2010). A common error is F0 doubling or halving errors (Hosom, 2005): For some segments, the algorithm decides on F0 values that are a fraction or a multiple of the actual value. These are due to happen when the allowed F0 range is not correctly set, and they can also happen for certain voice characteristics, such as vocal fry (Johnson, 2003, Section 3.3.3). These become apparent as jumps (discontinuities) at certain time points in the F0 curve.

We tried to ensure good quality F0 curves and through that reliable results using a multi-step process, as follows:

1. Set the parameters for an F0 tracker to suite the speaker characteristics.
2. Post-process it by automatically identifing and correcting F0 jumps.
3. Smooth the result by applying a median filter with a window size of 5, similarly to Ahmadi and Spanias (1999).

Below we summarize the method for setting the F0 tracker parameters and for dealing with the issue of F0 doubling and having errors.

3.1.1 Setting F0 tracker parameters

We extracted F0 contours using the Snack toolkit (Sjölander, 2006) and the ESPS method, which implements the RAPT algorithm (Talkin, 1995). Unless specified otherwise for the data used for a particular experiment, the window length was 20 ms and the shift was 10 ms, initially tracking F0 with a frequency range of 100-1200 Hz, which is suitable for children in general. Based on the initial result, we estimated the frequency range for the particular utterance based on the work of De Looze and Rauzy (2009), which was shown by Evanini et al. (2010) to optimize the performance of diverse F0 tracking algorithms to almost the same level:

$$\mathrm{F0}_{\mathrm{floor}} = \mathrm{quantile}_{35} \cdot 0.72 - 10$$

$$\mathrm{F0}_{\mathrm{ceiling}} = \mathrm{quantile}_{65} \cdot 1.9 + 10$$

Afterwards, we tracked the F0 curve again with this range and used this version in our analyses.

3.1.2 Project on automatic correction of F0 tracking errors

We automatically identified and corrected jumps in the outputs of F0 trackers. This approach was motivated by and resembles an unpublished solution created by Emily Tucker Prud'hommeaux while at CSLU; hereafter we do not deal with her work, and she was not involved in the development or the evaluation of the approach. We corrected the curves by multiplying or dividing the F0 values in certain segments so that they become continuous with neighboring segments. We made use of statistical features of all speech recordings for the speaker at hand when deciding about alternatives. This F0 pre-processing step improved the performance of our models significantly in several cases, for example for emotion recognition with Functional Principal Component Analysis features (see Section 8.2).

Note that this problem is complementary to the one solved by F0 trackers, which many people have invested years of research into (for example de Cheveigné and Kawahara, 2002; Hosom, 2005; Boersma and Weenink, 2009; Brookes, 2011), accordingly our methods are different. Whereas an F0 tracker generates an F0 curve for a speech wave form taking into account only one utterance at a time, our goal is to be able to improve the output of any F0 tracker, taking into account

all speech recordings for a speaker. Therefore we believe it is theoretically sound to perform F0 correction on a corpus as a secondary step after using a general-purpose F0 tracker.

Our F0 correction algorithm performs four tasks for each utterance: First, it smoothes the F0 curve with a median filter using a 5-frame window. Second, it identifies places where there may be a jump, indicating an F0 tracking error: If multiplying one of two neighboring voiced segments with an integer value would improve how well they fit our continuity criteria (see below), then it hypothesizes that there is tracker error. Third, for each jump it determines whether the segment to its left or right needs to be multiplied to correct the error and with what value. It decides which of the two consecutive segments is more reliable based on features of the segments: Segments that are shorter, have less energy, or are farther from the expected range of F0 values (calculated from all speech of the subject) are considered less reliable. Fourth, it applies only one of the determined changes, specifically the smallest modification on the original (unsmoothed) F0 curve. It repeats the whole process until there are no more jumps. Finally, a median filter with a window width of 5 is applied to smooth out single-frame discontinuities. This step gets rid of isolated unvoiced frames in a voiced segment and smoothes protruding F0 values to fit their neighbors.

The continuity criteria are the following:

- Two neighboring ends of two consecutive voiced frames should be close to each other in the sense that a jump does not explain the difference better.
- The tangents of two voiced segments separated by a short unvoiced segment (shorter than a threshold) should not enclose a high angle with the line connecting the two segments.

We set up a listening task, similar to that of McGonegal et al. (1977) to evaluate of the corrected F0 curves for the CSLU ERPA ADOS Corpus (see Section 4.2.1). This test helped to identify the most reliable F0 tracking method among the ones available to us at the time. Based on this, we chose to use the F0 tracker from the Snack toolkit and the above F0 correction algorithm.

We created the stimulus set like this: We detected F0 in the corpus using three F0 trackers, then corrected their output using a simple method and our correction algorithm. This way we had nine curves for each utterance: three curves for each of the three F0 trackers. We identified the utterances for which at least two of the nine curves differed grossly (difference $> 20\%$ on a length of at least 1 second), and assessed their quality through a listening experiment on a random subset of these utterances.

We conducted the listening experiment through Amazon Mechanical Turk (AMT). We asked AMT workers to listen to an utterance and to vocoded versions of the same utterance, synthesized with one of the detected F0 curves. We asked the raters to indicate if the melody of the synthesized

utterance was identical to that of the original one, if one was an octove higher than the other, or if they differed in some other way.

We collected three ratings for each utterance. At least two of the three labelers agreed 94% of the time. We chose to use the F0 tracker and correction method that got the most "identical" labels, namely Snack (Sjölander, 2006) and our correction method outlined above. At the time, we used the following settings for Snack: a frame shift of 10 ms, a frame size of 7.5 ms, which results in 133 Hz lower bound, and we set the upper bound to 600 Hz. This frame size was unfortunately incorrect, being shorter than the frame shift, as 25% of the data was not analyzed. Later we changed these settings as described above.

3.2 Computational models of prosody

Researchers have proposed several ways to model intonation, that is the fundamental frequency curve for utterances; for a review, see for example van Santen et al. (2008). One of the most prominent ways is the superpositional model, which we used in this work. This model assumes that the intonation contour can be quasi-additively decomposed into component curves. Examples of this model include the Fujisaki model (Fujisaki, 1981), the General Linear Alignment Model (van Santen and Möbius, 2000), and its simplified variant called SLAM (van Santen et al., 2004). None of them can claim to be perfect of course, especially because prosody is not yet completely understood (Peppé, 2009). We briefly review them below.

3.2.1 Fujisaki model

A prominent representative of the superpositional model is the Fujisaki model (Fujisaki, 1981; van Santen et al., 2004). It postulates that the intonation curve can be constructed by adding up (in the log-domain) a minimum value, a phrase curve, and accent curves. The curves are generated by applying filters to pulses: a filter to a Dirac pulse to generate the phrase curve, and a filter to rectangular pulses to generate the accent curves. A strong feature of this model is that it has relatively few parameters, which is crucial for parameter estimability. However, by the same token, this model may not be able to fit the great variety of pitch movements that can occur, in particular pitch movements that have a relatively fast rise and slow fall. In addition, again as a result of its simplicity, alignment with underlying segments is not specifically modeled, which may result in misalignment.

3.2.2 Generalized Linear Alignment Model

Another example of a superpositional model is the Generalized Linear Alignment Model (van Santen and Möbius, 2000), which posits that the component curves are a phrase curve, accent curves, and a segmental perturbation curve. The accent curves are non-linear time warps of an accent curve template. Alignment with the utterance segments is carefully modeled.

3.2.3 Simplified Linear Alignment Model

In its general form, the Linear Alignment model is an abstract conceptual model that is not amenable to efficient parameter estimation. The Simplified Linear Alignment Model (van Santen et al., 2004) introduces contraints to make the search for the optimal parameters feasible. In this model, the phrase curve is a linear interpolation between the phrase start, the syllable with the nuclear accent, and the phrase end points, and no segmental perturbation curve is used. The accent curves are created using cosine interpolation between the foot start, the F0 peak, and the foot end, where foot means an accented syllable together with the following unaccented ones. The model deals with three types of accents, initial, medial, and final, which are added relative to the intonation curve. The question and comma intonation components are also added relative to the phrase curve. The accent heights are model parameters. For an illustration of the model parameters, see Figure 3.1.

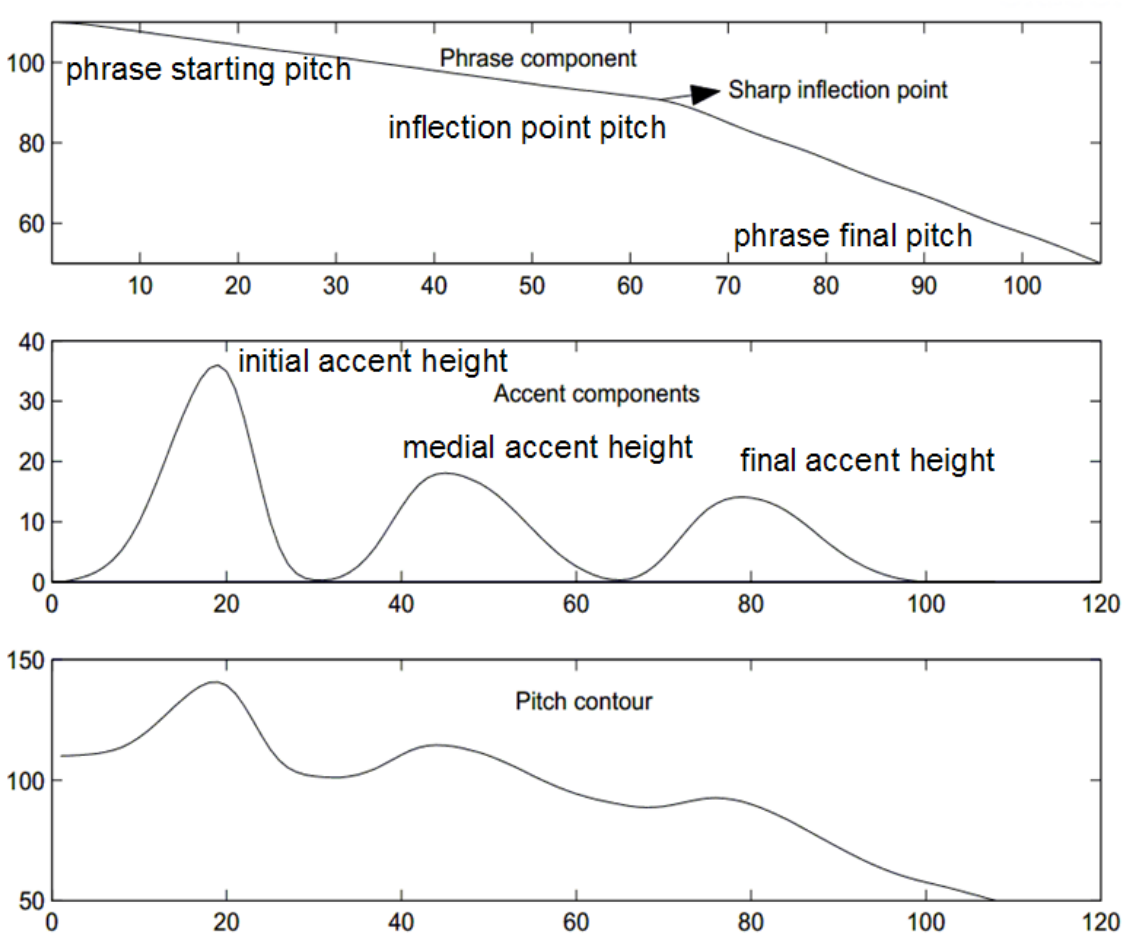


Figure 3.1: Illustration for the Simplified Linear Alignment Model parameters (from van Santen et al. (2004), with modifications)

3.3 Functional data analysis

Functional Data Analysis (FDA) is a method to analyze time series data as a set of continuous functions. To apply this methodology, one first needs to convert the data to functional form by fitting basis functions to the samples, such as B-splines, using a roughness penalty to avoid fitting the noise in the data. Once the data is in functional form, one can apply the functional versions of well-known statistical procedures, such as Principal Component Analysis (PCA). Below we summarize the method based on Ramsay et al. (2009); see also Gubian (2013) for its application in phonetic research.

3.3.1 FDA for prosody research

An intonation curve represented as a time series of F0 or intensity values can be treated as functional data, as we can consider it to be a sequence of noisy samples from a hypothetical continuous function. Accordingly, FDA has been successfully applied to various problems related to the analysis of speech prosody (see e.g. Gubian et al., 2010, 2011). For phonetic analyses, it is generally beneficial to align the curves at important landmarks (landmark registration; Ramsay et al., 2009), to separate amplitude and phase information and to avoid that important variation is extinguished. This may not always be possible, as it requires knowing the lexical content of each utterance, and the lexical content of aligned utterances needs to be comparable. When neither of these conditions is met, the method may still be applied without this step (see e.g. Arias et al. (2013), and our work in Chapters 6 and 8).

3.3.2 Converting intonation curves to a functional form

Consider the F0 or intensity curve for an utterance y consisting of n time points $t_j, j = 1 \ \ldots \ n$, and a set of K basis functions $\phi_k, k = 1 \ \ldots \ K$. We express y as a weighted sum of the basis functions to capture the most important details of the curve:

$$\hat{y} = \sum_{k=1}^{K} c_k \cdot \phi_k.$$

Instead of a mathematical representation of the basis functions, it suffices to know their values at the time points t_j, which we represent as a matrix

$$\Phi = \Big[\phi_k(t_j)\Big].$$

We want to choose the weights c_k above to minimize the sum of squares errors between the original y and its approximation $\hat{y}$. We use the least squares error function, assuming Gaussian

noise on y:

$$SSE(\hat{y}) = \sum_{j=1}^{n} (y(t_j) - \hat{y}(t_j))^2 = \Phi^T \cdot c.$$

We refine the solution by adding to the error function the integral of the second derivative of $\hat{y}(t)$ with respect to t as a roughness penalty, to avoid close fit to the noise. This gives us the regularized least squares solution, with the following estimate for the coefficient vector:

$$\hat{c} = (\Phi^T \Phi + \lambda R)^{-1} \Phi^T y \quad \text{where}$$
$$R = \int D^2\Phi(t) \cdot D^2\Phi^T(t) dt.$$

3.3.3 Functional Principal Component Analysis

For a set of N curves, *Functional Principal Component Analysis* (fPCA) identifies the first M orthogonal eigenfunctions that capture most variation. We represent curve i as a weighted sum of eigenfunctions ξ_m, with a set of coefficients that maximizes

$$\sum_{i=1}^{N} \int \xi_m(t) \cdot \hat{y}_i(t) dt$$

subject to

$$\int \xi_m^2(t) dt = 1 \text{ and } \int \xi_m(t) \cdot \xi_l(t) dt = 0, \forall l < m.$$

That is, we look for a set of normalized orthogonal eigenfunctions that can explain most variance in the set of curves.

The fPCA coefficients are given by

$$c_{i,m} = \int \xi_m(t) \cdot (\hat{y}_i(t) - \bar{\hat{y}}(t)) \; dt, i = 1 \; \ldots \; N, m = 1 \; \ldots \; M$$

for utterance i, eigenfunction m, and mean curve $\bar{\hat{y}}(t)$.

When we work with a speech corpus, we can analyze all utterance curves together regardless of their labels. During the analyses, we can work with the fPCA coefficients $c_{i,m}$ for the utterances. For example, to identify group differences, we can compare the first n coefficients between the utterances for each group. Or we can visualize the component curves associated with certain continuous features.

3.4 Statistical Techniques

In this work, we generally deal with hierarchical data, for example multiple utterances from each child in a subject pool. To satisfy the independence assumption (see e.g. Winter, 2011), we work

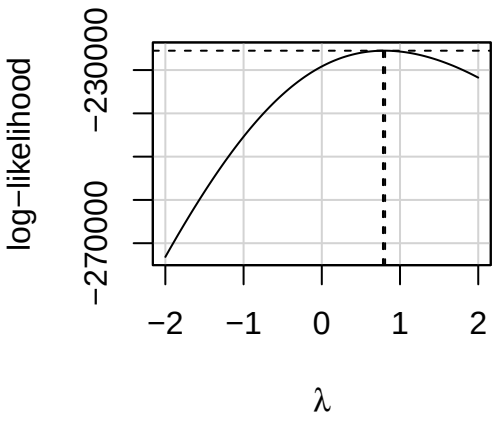


Figure 3.2: Illustration for the Box–Cox transformation for finding the optimal Yeo–Johnson power transformation for speaking rate (in syllables per second). The best power for this example is around 0.8. Generated by the `car` R package.

with per-group summary statistics, with generalized mixed effect models, or appropriately Monte Carlo tests that treat hierarchical data correctly (see below).

For linear models, we always make sure that the model assumptions are not violated, namely that of linear relationship between the dependent variable and the predictors (also called independent or explanatory variables, predictors, or regressors), homoscedasticity, and the normality of the residuals. We use power transformations to correct issues with these: We make the distribution of the dependent variable approximately normal using the Box–Cox transformation (Box and Cox, 1964; for an illustration, see Figure 3.2) and the Yeo–Johnson family of modified power transformations (Weisberg, 2001):

$$\hat{y} = \frac{(y+1)^P - 1}{P}.$$

For the predictors, one can decide on a power transformation using Tukey's bulging rule (Hoaglin, 2003). One can also use other transformations to handle collinearity of predictors, such as standardization, or polynomials of the predictors. Plots of the predictors against the dependent variable help to see if the relationship between them is approximately linear, or if some of the predictors need a power transformation.

Unusual and influentual data points require special attention, as a few outliers can have a large effect on the model parameters. Among other things, visualization techniques can help to identify such points in the data set, including boxplots that show outliers and residual plots. Sometimes

we address this issue simply by using robust statistical features.

3.4.1 False Discovery Rate correction

When testing multiple hypotheses, we either do planned comparisons, or if doing exploratory analyses, we use False Discovery Rate correction (FDR; Benjamini and Hochberg, 1995; Klaus and Strimmer, 2013) to reduce the chance of finding spurious relationships. When we find a significant difference among multiple groups, we do post-hoc testing using Tukey's HSD.

3.4.2 Mixed effect linear models

Mixed effect linear models (MELM) can explain the relationship between the dependent variable and multiple possible explanatory variables taking into account random factors as well (such as the subject's identity for a set of utterances for multiple subjects), if the prerequisites of linear models are granted. The variable transformations described earlier are often effective in removing non-linearity, multi-collinearity, and heteroscedasticity. It is important to include in the model the relevant predictors but not to complicate the model beyond what is necessary, both to make it easier to understand and to ensure that enough data is available for estimating the model parameters with sufficient precision.

We apply such models to our data using R and the `lme4` package (R Core Team, 2017; Bates et al., 2015), adhering to the following protocol: We start out with a model that contains all explanatory variables that may have an effect on the outcome, both as fixed effects and as slopes for the random variable(s) besides the random variable intercept. After transforming the variables if necessary as outlined above (to eliminate heteroscedasticity of the outcome, collinearity among the predictors, etc.), we remove the non-significant effects by an exhaustive search over all sub-models of the full model to find the model with the highest Akaike Information Criterion using the `MuMIn` R package (Bartoń, 2016), or backward ellimination based on the likelihood ratio test using the `lmerTest` R package (Kuznetsova et al., 2016). We report the results for this minimal model after making sure that the residuals do not show pronounced heteroscedasticity or deviation from normality.

The aspects of this model and its fit to the data that we report about are:

- the dependent variable;
- the fixed effects (other than those with negligible effect sizes) and their interactions;
- the random effects together with the fixed effects for which slopes are calculated;

- whether each variable is continuous or categorical;
- the marginal and conditional coefficient of determination, which are the variance explained by the fixed effects and both the fixed and the random effects together, respectively (using the MuMIn R package by Bartoń, 2016);
- the coefficients for fixed effects, possibly with their confidence intervals if these are wide;
- the p-values for the fixed effects that are significantly different from the intercept after FDR correction; the p-values are calculated based on Satterthwaite's approximations using the `lmerTest` R package (Kuznetsova et al., 2016);
- all pairwise group differences using a post-hoc test with Tukey's HSD.

3.4.3 Monte Carlo significance test

When some of the assumptions of linear models may not be met, or if we want to compare a value between groups without controlling for potential confounders (other than by matching the data set on them), we generally use Monte Carlo testing: We randomly shuffle the group membership of the samples a number of times (e.g. 10,000) and calculate the ratio of cases that have a test statistic that is more extreme than the one belonging to the actual group membership distribution. As the test statistic, we use the F-statistic. For hierarchical data (for example when there are multiple utterances for each of several subjects) we make sure that the group membership of samples belonging to the same group always get the same random group membership.

When applied to multiple variables, it is similar to doing a MANOVA, only we do not need to make sure the assumptions associated with the MANOVA are satisfied. Another method we utilize is 50–50 Manova by Langsrud (2005): It does not handle random effects, but handles multiple collinear predictors. We use the `ffmanova` implementation (Langsrud and Mevik, 2012), which can calculate adjusted p-values using Monte Carlo simulations according to familywise error rates and FDR correction.

3.4.4 Reliability of results

We do not do a priori power analyses because in our case the sample sizes were pre-defined in the data available for our analyses, moreover, it is often hard to know the expected effect size. We do not do retrospective power analyes either as it is not recommended in general (Hoenig and Heisey, 2001). Instead, we report 95% confidence intervals to show how important a particular result is. We report effect sizes using Cohen's kappa, or Pearson's r (product-moment) correlations

between the target and the predicted values. For MELM models, we report the coefficients and their confidence intervals.

Chapter 4

Corpora

In this chapter, we describe the speech corpora used as the basis for our work. We give details about the kind and amount of data contained in it, the extra information we derived from that, and how we turned it into a format that yields itself to statistical analyses. This is important for interpreting our results, as we need the right type of data for answering our questions, and the quantity of the available data limits the statistical power of our tests. The data also need to meet certain requirements, namely, ecological validity, external validity, and internal validity. It is ecological valid if the procedure resembles the real world situation. It is externally valid to the extent that the results generalize to the real world. It is internally valid as much as a causal conclusion based on a study is warranted; we can attain this by minimizing the systematic bias, that is, not biasing the data collection in a way that would boost the chances of a certain finding. This includes using standardized instructions, avoiding investigator effects (when that examiners influences subject behavior in a particular direction) and demand characteristics (when the subjects change their behavior to a perceived expection), counterbalancing (presenting examples randomly to avoid systematic influence of stimulus order), and controlling for potential confounders. Below we examine our corpus in the light of these requirements.

4.1 CSLU Cross-modal Corpus

This corpus was created by colleagues at the Center for Spoken Language Understanding (CSLU). We briefly summarize it based on van Santen (2014) and Asgari et al. (2014). The subjects were 28 children aged 8–11: 10 with TD, 11 with ASD, 4 with SLI, 3 with Unspecified Developmental Delay. They were trained to re-enact a brief story. When a child spoke, his or her whole figure was recorded on video. Then from each video, five different artifacts were derived: the speech recording, the speech rendered unintelligible (delexicalized speech; Kain and van Santen, 2010), the textual transcript, a video containing only the child's face, and the video with the face blanked out (gestures).

The final data set contained 19–46 sentences per subject (with a median of 27), with 835 sentences in total for each of the five modalities. Ten raters estimated the arousal and valence for each stimulus on a 1–5 scale; each utterance was assessed by each rater. An aggregate score was derived by averaging the ratings. Jan van Santen assessed the reliability of the subjective ratings by randomly splitting the ten raters into two groups 100 times, and calculating and averaging the correlation between the aggregated ratings derived from the two groups of raters. The results are included in Table 4.1.

Table 4.1: Average correlation between the aggregated emotion ratings for randomly selected subgroups of raters for the CSLU Cross-modal Corpus

Modality	Arousal	Valence
Text	0.75	0.93
Speech	0.88	0.91
Delexicalized Speech	0.85	0.75

4.2 CSLU ERPA Autism Corpus

We worked with a relatively large and well-characterized corpus, the CSLU ERPA Autism Corpus. The corpus contains the speech recordings for various standardized tasks that are used either for diagnosing autism or for measuring cognitive and language abilities. These are recordings of the Neuropsychological Assessment (NEPSY), Non-word Repetition Task (NRT), Clinical Evaluation of Language Fundamentals (CELF), Verbal Fluency, and Autism Diagnostic Observation Schedule (ADOS) sessions. The tasks were administered by trained psychologists, who also rated the children's performance on these tasks. It is not our duty here to go into detail about these tasks, except for the ADOS, as we worked with this part of the corpus.

4.2.1 The Autism Diagnostic Observation Schedule (ADOS)

The Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000; Gotham et al., 2007) is a session of interactions between an examiner and a child, about an hour long, consisting of a specific sequence of activities and discussions that were selected to bring out behaviors that are typical of autism. It comprises four activities, denoted later on as Play, Picture description, and Wordless Picture Book, and Conversation. The ADOS tasks corresponding to the first three are: "Make Believe Play & Joint Interactive Play"; "Description of a Picture"; and "Telling a Story from a Book". Their meaning is obvious, except maybe for the last one, wherein the child was asked to narrate a wordless picture book. In the Conversation part, examiners must talk about four particular topics (and even how they should start and maintain the conversation is prescribed), belonging to the following ADOS activities: "Emotions"; "Social Difficulties and Annoyance"; "Friends, Relationships and Marriage"; and "Loneliness". We also assigned the conversations initiated during the Picture Description task to the Conversation activity. The examiner needs to rate the child's behavior on several scales, including one for prosody, but the prosody ratings are

not used in the algorithm for diagnosing the child, probably due to high variability in the examiner scores (Peppé et al., 2011, p.51).

4.2.2 Data collection and corpus characteristics

The CSLU ERPA Autism Corpus was collected at the Center for Spoken Language Understanding (CSLU) between 2005 and 2012, in the course of a large NIH-supported project on expressive and receptive prosody in autism. It contains videotaped sessions of child–examiner interactions, and manual transcriptions of the speech content as text, annotated using the Systematic Analysis of Language Transcripts (SALT notation; Miller and Chapman, 1985). The subjects were 241 children, native monolingual speakers of American English aged 4 to 9, of whom eventually 113 received a Best Estimate Clinical (BEC) consensus judgment and became a part of the study.

Inclusion and exclusion criteria included the following (Hill et al., 2015a): The children must not have a known metabolic, neurological, or genetic disorder other than what pertains to their diagnoses, no sensory motor impairment, brain lesions, or orofacial abnormality. Their mean length of utterance in morphemes has to be at least three. They must not have ID (non-verbal IQ at least 80). They must not have speech intelligibility impairments. See more details on the cognitive measures in Hill et al. (2015a).

4.2.3 Diagnostic categories

The subjects were high-functioning children with ASD (abbreviated as HFA) with language impairment (ALI) or with normal language (ALN), and the control groups were formed by children with Specific Language Impairment (SLI) or typical development (TD). Selecting the control groups like this can help us to separate the differences that autism causes per se from the effect of language impairment. This can help to explain some of the heterogeneity in the ASD population, as well as to see whether a certain symptom is specific to autism or not. The children in the SLI group scored in the normal range on all measures of autism-related symptomatology, while the HFA groups met all conventional cut-offs for HFA. We summarized the corpus characteristics in Table 4.2.

Table 4.2: Summary statistics for the CSLU ERPA ADOS Corpus. The cells contain either a count or the mean and the standard deviation of values for the subjects.

DX group	subject count	age (years:months)	speech duration seconds	sentence count
TD	43	6;3 (1;3)	921 (356)	437 (150)
ALN	25	6;5 (1;4)	851 (512)	397 (203)
ALI	26	6;9 (1;1)	682 (413)	359 (178)
SLI	19	7;1 (1;0)	939 (453)	507 (189)

One of the strengths of the CSLU ERPA Autism Speech Corpus is that the children were selected based on well-defined criteria, and are very well characterized. The diagnoses were determined in extensive clinical consensus meetings in accordance with the DSM-IV (American Psychiatric Association and others, 2000) criteria and the published cut-off scores on the ADOS (Lord et al., 2000) and the Social Communication Questionnaire (SCQ; Rutter et al., 2007). The selection criteria for SLI included a documented history of language delay or language deficits, CELF scores with one standard deviation below the mean, and a BEC judgement of language impairment but no ASD (Hill et al., 2015a). As a result, at least 30% of children recruited for this group were excluded. Most importantly, this process ensured that the children in the SLI group were quite different from the HFA group in terms of ASD features, even though their ADOS severity score is somewhat higher than that of the TD group, as can be seen in Table 4.3 below.

4.2.4 Matched groups

In our analyses, we matched pairs of groups on relevant measures such as age, IQ, and autism severity, making sure that they do not differ significantly on those. Our goals were threefold: First, to match pairs of groups on measures in such a way that they will be similar on potential confounding variables, while not changing the group setups substantially (see more on that below). Second, to keep as many subjects as possible, both to preserve power to detect potential differences and to avoid getting rid of variation inherent in the groups being studied. Third, to minimize the difference between the groups beyond ensuring that they are matched.

All groups matched together

We follow established research practices on our matching criteria. Papers that use group-matching generally match the means of the distributions (see e.g. Rubin, 1973). Facon et al. (2011) also

show that it is advisable to match the overall shape of the distributions as well. They used the Kolmogorov–Smirnov (KS) test for that, but it is not suitable for us, as KS is for continuous distributions, whereas some of our variables are not (e.g. ordinal values from the 1–10 range). The Anderson–Darling (AD) test can also be used to test if two samples come from the same distribution and it can be used for discrete-valued variables as well; moreover e.g. Razali and Wah (2011) has shown AD to have higher power than KS for smaller sample sizes, and the number of our subjects can be considered small with their standard. Regarding the minimum significance level α to use, Mervis and Klein-Tasman (2004) (as well as Facon et al., 2011) proposed using $p > 0.50$, also mentioning $p > 0.20$ as values to consider.

For our data, we ensured that both the means and the overall shape of the distributions of the covariates are similar between the groups, defined as having $p > 0.20$ two-tailed on both the t-test and the AD-test. We did not use a higher lower bound for p so as not to reduce the subject pool too much. Regarding the matching procedure and the implementation of these test used, see Chapter 5, especially Section 5.4.

We came up with one set of matched subjects from each diagnostic group (i.e., TD, ALN, ALI, and SLI) by eliminating some subjects from the pool; see Table 4.3. We made sure that all four groups are matched on age. We also made sure that pairs of these groups are matched on relevant measures: We matched TD and ALN as well as ALI and SLI on verbal IQ (VIQ) and performance IQ (PIQ; also called non-verbal IQ: NVIQ), and ALI and ALN on their autism severity score. For VIQ and PIQ, we used the Concatenated WPPSI3 and WISC4 Standard Scores (Saklofske et al., 2003), and for autism severity, the ADOS Calibrated Severity Score (Gotham et al., 2009). We kept all subjects with SLI, as that is the group with the least number of subjects, striving to keep as many as possible of the ALI, ALN, and TD groups, in this order of priority.

The rational for the above criteria is that the TD and ALN groups are naturally very close on both problem solving and language measures, as the ALN group has normal intelligence and normal language, the ALI and SLI groups are similar on the same because members of both groups have language impairment, and the ALN and ALI groups have similar autism severity rates, despite differing on intelligence and language measures. Had we matched all four groups on all measures, we would have lost all children on the lower end of the scales from the disordered groups, and all TD children from the higher end of the scales, distorting the original setup of these groups and thus decreasing the credibility of our results.

Table 4.3: Summary statistics for the CSLU ERPA Corpus groups matched on age, VIQ, PIQ, and the ADOS score. The cells contain the mean and standard deviation of the values for the groups. VIQ and PIQ are Concatenated WPPSI3 and WISC4 Standard Scores, autism severity is the ADOS Calibrated Severity Score.

DX	subject count	age (years:months)	VIQ	PIQ	autism severity
TD	31	6;9 (1;0)	115 (11)	116 (13)	1.2 (0.5)
ALN	19	6;8 (1;2)	110 (14)	117 (17)	7.2 (1.9)
ALI	25	6;8 (1;2)	83 (8)	104 (17)	7.8 (1.8)
SLI	19	7;1 (1;0)	86 (6)	102 (12)	2.9 (2.6)

Pairwise matched groups

For some of our early analyses, we matched pairs of groups separately from each other. Thus the subjects for a group may be different in different comparisons; for example the TD subjects in the TD–ALN pair are different from those in the TD–SLI pair. We report such analyses, possibly also redoing them using the later matching configuration shown above, because we have published some of our findings using this setup.

The subject group pairs are the following:

- TD (19) vs. ALN (18) matched on age, VIQ, and PIQ
- TD (22) vs. SLI (17) matched on age
- SLI (17) vs. ALI (18) matched on age, VIQ, and PIQ
- ALI (18) vs. ALN (20) matched on age

We used a version of the `heuristic2` algorithm described in Section 5.3 to create the matched groups.

4.2.5 ADOS recordings

For this study, we analyzed ADOS audio recordings (see Section 4.2.1 above). An advantage of using ADOS sessions is that these are quasi-natural conversations, whereas the topics (and even the toys the children play with) are standardized. They are of course different from a family conversation, but being recorded "in an unfamiliar, but positive, fairly relaxed, setting" (Lord, 2010), and not being specialized tasks with pre-determined content like some other diagnostic intruments, they can be regarded as spontaneous speech,

We believe this ADOS data meets the requirements of ecological and external validity, as well as internal validity: The ADOS session resembles the everyday situation when a caretaker or teacher sits down with the child to talk and to play. The clinicians were trained to administer it to children, were blind to the children's earlier diagnostic statuses, and followed standardized instructions. Moreover, since the children are well characterized, we can control for potential confounders in our analyses.

Transcription

The speech has been both segmented and transcribed, and the transcripts aligned to the speech waveforms. The speech is divided into so-called communication units or C-units ("an independent clause with its modifiers"; Loban, 1976). We made use of the textual transcriptions and the SALT annotations to identify utterance features, therefore we briefly review as much of this transcription standard as is relevant to our work.

SALT defines the sentence-final punctuation marks to use for marking:

- statements (denoted throughout this work by "S"),
- exclamations ("E"),
- questions, including yes–no questions ("YN"), WH-questions ("WH"), and open-ended utterances ("O"), which are used to prompt the other speaker to finish the sentences
- as well as interrupted ("I") and abandoned ("A") utterances.

Since children seldom used open-ended utterances, we do not deal with those, but systematically analyze the others. The sentence-internal and sentence-final punctuation marks are used consistently throughout the corpus, and thus determine one of our analysis units, the sentence.

We also made use of the SALT markup that labels speech phenomena to filter out utterances that we could have caused issues in our analyses. Unless noted otherwise, we excluded utterances containing:

- unintelligible segments (e.g. "X" for an unintelligible word),
- non-speech phenomena (e.g. a cough denoted as ".cough"),
- mazes (mazes are revised or abandoned parts of the utterances, e.g. the words in parentheses in the following sentence: "I (want to) (uh) need to go to (be*) bed now."),
- grammatical errors (e.g. "He *is go/ing home" where the child omitted the word "is"), and

- non-original speech productions, such as scripted speech, (immediate) echolalia, palilalia, and perseveration.

4.2.6 Relational Feature Tables from ADOS Corpora

The ADOS recordings are basically unstructured data, but we can nevertheless store it as *tidy data* (see e.g. Wickham, 2014) in relational database tables, which yields itself to statistical analysis more easily. Our basic unit of analysis is the subject, and for each subject, a number of sentences. Sentences in turn are made up of C-units, those of words, and those of phonemes. A speech utterance can be regarded as a sequence of overlapping frames for signal processing (e.g. F0 tracking; see Section 3.1). We created a table for each of these levels and one record for each unit within a level, this storing all information that is relevant for us in a structured way.

4.2.7 Determining sentence type

One piece of information necessary for our analyses was the type of each sentence in the corpus. We assumed that transcribers who listened to the conversation correctly determined for each utterance if it is a question or a statement. When doing this, we are not relying on the abilities of the children, as we cannot be sure that young children and the children with atypical development use prosody effectively to communicate the question–statement distinction. Rather, we trust the transcribers, who listened to the utterances in context.

The transcription guidelines for the CSLU ERPA ADOS Corpus directed the transcribers to

> Mark everything a question that plays the role of a question, even if its grammatical structure indicates otherwise.

In other words, the transcribers were to mark the interpretation of the examiner, who interpreted the child's intentions and responded in a certain way, determining the role the utterance in the dialog. So if the examiner treated an utterance of the child as a question and the child seemed to accept that interpretation, then we expect that the transcriber put a question mark at the end of the child's sentence. Unless the child readily gives up his original intention when misinterpreted, it should be obvious from the full discourse what his or her intention was. The sentence type decision is up to the interpretation of the transcriber only when it is not obvious from the examiner's response if she interpreted it as a question or a statement.

For the questions, we determined automatically the sentence-subtype, especially whether a question is a yes–no question (more formally "polar" question, denoted hereafter as YN-question) or a WH-question. For this, we first review what sentence types exist and then outline our approach

for deciding on these automatically. We thank Masoud Rouhizadeh for showing his approach to this problem, which helped us refine our approach by comparing the results of the two.

On sentence types

We distinguish multiple sentence types, including statement (in writing, it ends in a full stop), exclamation and imperative (both end in an exclamation mark), and multiple question types (all end in a question mark). One may tend to view the distinction of sentence types as a simple matter, yet it can get quite complicated if one tries to do that by analyzing the surface form. We need to distinguish different question types for our work because they have different associated prosody, although the same question can occasionally be said with multiple very different intonations as we shall see.

YN-questions have a rising intonation at the end in English (and in many other languages, but not e.g. in Hungarian). A statement can be turned into a YN-question by adding a tag-question; for example "You remembered the eggs, right?", or "didn't you?" The latter can be said with a statement intonation. Rhetorical questions also have no change in intonation. The question mark is sometimes omitted when there is no question intonation; for example "It's too late, isn't it." Conversely, we can say a statement with YN-question intonation, turning it into a question, for example "You're going?"

A special kind of polar question is a choice question (also called disjunctive question or alternative question). For example, "Do you like tea or coffee?" It can also be interpreted as a YN-question: "you like one of the two, or you don't", and thus can be answered by either "yes" or "no".

The so-called "WH-questions" ask for new information using the words "which", "what", "who", "whom", "whose", "when", "where", "why", and "how" (all containing the "wh" letter combination, except for "how"), and have very similar intonation to statements except that the question word is generally quite emphatic. They usually have inversion (i.e the subject and verb exchange places), as in "What did you do?". But WH-question can occur with no inversion as well; for example "You did what?". Interestingly, certain statements can be considered as indirect questions; for example "I wonder where Jack is" meaning "Where is Jack?".

The SALT guidelines use three more sentence-final punctuation marks, which however are not relevant to our study: Open-ended questions ask for information by starting the answer itself. This type of question hardly ever occurs for the children that we worked with. Abandoned and interrupted utterances are differentiated as well. We generally did not use sentences of these types, and it would obviously not always be possible to infer these based on the sentence text alone either.

For these reasons, we did not classify utterances into these sentence types automatically, but relied on the transcriber's decision.

Some frequent sentence structures

Different sentence types correspond to different sentence structures, with some overlap between the categories. Here we describe some frequent and characteristic ones, as well as some that are easily confusable (at least algorithmically) , and some ambiguous ones.

The typical WH-question contains "<WH-word> <auxiliary>", such as "When is ...?" The WH-words again are: "when", "where", "why", "whom", "who", as well as "what", "which", "how", and "whose". The first five words should have an auxiliary right after them in a WH-question, whereas the last four can be followed by various expressions, such as "What kind of animals do you like?" and "How many apples are there?" Moreover, "who", "what", and "which" can be followed by an auxiliary even when they are not question words but conjunctions, playing the role of subject in a subordinate clause. For example, "And there was a turtle, on a log, who was look/ing around." So these can be present and followed by an auxiliary in statements and YN-questions as well, which shows that a deeper analysis of the structure may be necessary to determine the sentence type.

Note that in a question-clause, we must have a verb somewhere after the subject ("<auxiliary> <subject> ... <verb>"), unless it is a question tag (such as "isn't it?"). If there is no main verb, then the first verb must be the main verb and not an auxiliary (possibly in an imperative, or a sentence fragment where the subject was omitted). For example: "Have some ice cream."

The last word of a sentence can turn it into a WH-question irrespective of the structure of the first part. For example: "Do you think it is red, or what?", "There were two dogs, and a what?", and "Do what?" are WH-questions even though the first part in this example sentences is a YN-question, a statement, and a command, respectively.

YN-questions typically start with an "<auxiliary> <pronoun>" sequence, such as "Is it ...?", "Does he ...?", "Weren't they ...?". This part is frequently preceded by conjunctions as well as acknowledgments and other words. For example, "But is it ...?"; "Yes, and does he ...?"; "Hey, are you ...?" There can of course be any noun phrase instead of the pronoun, as in "Do the tasks please you?" (whereas "Do the tasks please!" is an imperative).

A question containing a WH-word can still be a YN-question if the WH-word is a conjunction introducing a subordinate clause. For example: "You mean Tom, who's twenty-four, has a baby?" The following utterance is also a common example of this phenomenon in the CSLU ERPA ADOS corpus: "And so when they did that to you, do you think you would say something?" In this case

the starting "when" refers to the time of the event and is not a question-word. It is not to be confused with a relatively similar WH-question format: "And so when did they do that to you , do you think?" As we can see, the "do you think" clause can be part of either a YN-question or a WH-question, depending on whether the clause containing the WH-word has a typical or inverse word order. A sentence can also have a clear-cut WH-question format, and yet it may be a YN-question for pragmatic reasons: "What made me mad?" Here it is apparent that the speaker is not likely to be asking this about himself or herself, but is rather asking if this is what the conversation partner would like to know.

The guidelines also state that an utterance is to be considered a question "even if the question is in the form of a quote. C: He said, is that you, frog?" In practice, this also means that when a question consists of multiple parts, it is the last part that dominates. For example, when a yes–no and a WH-question occur side by side, we classify it as having the sentence type of the last part (in the the following case, a WH-question): "Do you go to a swimming pool or where are you doing it?" Similarly, when the sentence ends in a statement, we consider it a statement: "<Oh, do you think you can do it, Dad?> is what the son said."

A statement-like structure can also play the role of a YN-question (carrying a question prosody). For example: "You really went there?"

For statements, even though they seem to be the simplest to identify, it is always the context that determines their role. As we saw in our previous example, basically any statement can be a YN-question as well. Moreover, a sentence that has a question structure may carry statement prosody. For example: "Aren't they nice."; "And probably that hurt your feelings, didn't it." These usually do not carry question prosody, but are not clearly statements either.

Also, one can often omit the subject of the sentence in casual speech. When it happens, the sentence starts with a verb, or an auxiliary immediately followed by the main verb. These look similar to imperatives, or even YN-questions, but they are not. For example: "Must flap my wings faster!"

Some examples for confusing sentence structures: "So how, for example, does a friend differ from a spouse?" Note that even though there is a clause starting with "<auxiliary> <noun phrase>", it is a WH-question, whereas "Help me, is what Riddick says." has a similar clause and is a statement.

Sometimes parentheticals make part of the sentence seem similar to a question, depending on where the clause boundary happens to be after inserting them. For example: "What makes him happy, I'm going to guess, is playing." In this case, the first part looks like a WH-question, but it is actually part of a statement.

Determining sentence type automatically

We determined the sentence type for the child utterances using a simple rule-based approach utilizing parse trees for the utterances and also doing some manual corrections. The task turned out to be simple in the majority of cases, but very hard for a relatively small fraction of the utterances.

We parsed the sentences with the BUBS Parser (Bodenstab and Dunlop, 2011) with the Berkeley SM6 latent-variable grammar (Petrov and Klein, 2007) that is provided with the parser. To simplify the task for the parser, we worked with the *intended text* of the utterances, which is a clean version of the utterance created by automatically removing mazes and errors based on the SALT markup provided by the transcribers. Even so, the parser could not parse some sentences, especially the sentence fragments, and came up with bad parses for some sentences made up of multiple clauses or those that are incorrectly segmented (e.g. when mazes were not marked correctly).

We relied on the transcription regarding whether a sentence is a statement or a question. We classified sentences ending in a question mark into two basic question categories outlined above: WH-questions (denoted WH), and YN-questions (denoted YN). The algorithm works the following way: We decide on YN when it does not contain a WH-word, or when there is one, but the question ends in asking for an acknowledgment (such as "OK?", or "didn't you?"), a single word request (e.g. "see?", "remember?"), or something that give away it was a request (e.g. "please?"). We do the same if it looks like a subordinate clause (i.e., "[[<auxiliary>] <subject>] <verb> .. <WH-word>"). We explicitly list frequent verbs that can take that place, such as "remember" and "know". Some words that are often used for asking for acknowledgment in the ADOS corpus are: "OK", "kay", "huh", "right", "alright", "really". Otherwise, when the parse is not available, we decide on WH when there is a WH-word in it. When the parse is available, we use a simple heuristic algorithm that decides between YN and WH based on the presence of certain parse tags.

Evaluation

We currently do not have a corpus with gold-standard sentence-type labels, therefore we needed to use approximate solutions. The author of this dissertation randomly checked some of the sentences, reading the sentence, determining its type, and comparing it to the automatically determined label. The manual and automatic labels agreed over 98% of the time. However high this number seems to be, it may still not be enough for applications where we expect every single sentence to be labeled correct, as was necessary for some of this work (see Section 7.2.5). In such cases, the only possibility is to review all labels.

Research directions

Applying machine learning One can use a machine learning algorithm to create a more systematic approach to the problem. The training and test set can be derived from the above initial sentence type labels. Possible features are n-grams of words, part-of-speech tags, and possibly parse tree node labels. One must make sure that the model does not overfit the training set so as to avoid creating a complex model that fits the labeling errors as well. It would benefit the model if someone can review the utterances for which the trained model predicts something else than the current sentence type, manually correct those that are wrong, and iterate the process as long as the labels improve.

We did not follow this approach because based on manual inspection, the sentence types determined by the above simpler approach seem to be correct in the majority of cases, and this topic was not the main focus of our research.

Labelling sentence types based on text alone Our goal is to examine prosody in relation to the sentence type, therefore it is desirable that we determine sentence type without taking into account the prosody, otherwise we may misinterpret the child's intentions if the child used prosody atypically. The transcribers were instructed to rely on the prosody only when determining the phrase boundaries and which parts belong to mazes. Yet the examiner, whose interpretations the transcribers were to record, probably took into account the prosody of the utterance, besides its structure and contents, when deciding.

We can ask labelers to look at the text of a set of child utterances in context and to choose the possible sentence types based on the text alone. This should happen by showing the preceding context only (namely the examiner's previous sentence or turn) and the child's current turn containing the sentence in question. The labeler is to choose the sentence types that are possible. The most interesting sentences to ask labels for are the ones that are ambiguous between several sentence types based on the text.

We expect that the labelers would choose the sentence type that we determined automatically from the sentence-final punctuation and the sentence text, and sometimes another one as well. For example, the sentence "You went there" might be a YN-question (asking for a confirmation), or a statement, depending on the context.

Exclamations

The transcribers were expected to mark exclamations, besides open-ended utterances (intended to prompt the other speaker by using a rising intonation and leaving off the last word), abandoned and

interrupted utterances, and of course statements and questions. Intuitively, whether an utterance should be considered a statement or an exclamation is mainly determined by the prosody. This is underlined by the fact that the text for over half of the exclamations also occured as statements. There are also sentences that end with an exclamation mark, but contain only pause fillers or sound effects, in which case it is obvious that the transcriber must have decided on using an exclamation mark based on the prosody.

Although it is not possible to determine with certainty if a sentence is an exclamation or a statement based on text alone, one would be able estimate its probability. The reason is that the likelihood of some content features seem to differ between statements and exclamations. For example, the ratio of mazes seems to be smaller among exclamations (about 6% vs. 13% among statements in the CSLU ERPA ADOS Corpus) and there are about twice as many imperatives in exclamations (about 5% vs. 2.5% for statements). Since deciding on whether an utterance is an exclamation is totally up to the interpretation of the transcriber, we decided to exclude exclamations from our analyses.

Chapter 5

Matching Diagnostic Groups

Having reviewed the necessary background, the computational techniques, and the speech corpora available for our analyses, we are now ready to start discussing our own contributions. In this chapter, we describe a problem whose solution is a preliminary step to our later analyses of the autism data. We deal with it theoretically and then describe a general computational solution, building upon earlier work by van Santen et al. (2010). Its application to our corpora was important for our present study, and it has been used by others at CSLU as well (see e.g. MacFarlane et al., 2017). We based this text partly on a paper we are preparing with co-authors Kyle Gorman and Jan van Santen (Kiss et al., 2017).

5.1 Motivation for Matching Subject Groups

When studying the effect of group differences on target variables, one needs to minimize the effect of confounding variables, which may influence both the dependent and the analyzed independent variables, so as not to draw false inferences. One can deal with potential confounding covariates at different stages: First, one can design the study such that one reduces the effect of confounders (e.g. by doing randomized controlled trials or using stratified sampling). Such designs can also take care of unobserved factors if the sample size is large enough and the selection is not biased. Second, one can use *matching*, that is, choose a subset of the subjects after the data collection that are similar in a certain way across the groups (Szatmari et al., 2004). Third, covariates can be dealt with in the analysis phase using statistical techniques that can take their effect into account (e.g. ANCOVA or multivariate regression). Szatmari et al. (2004, p.55) summarizes these strategies like this: "Stratification can be used when target and comparison participants can be divided into subgroups based on a small number of categorical levels of the confounder. Matching occurs when target and comparison participants are selected to be similar on the confounding variable, which is a more statistically powerful strategy. Multivariate analyses can be used when there are multiple confounding variables and when it is not possible to identify perfectly matched pairs." In this chapter, we motivate the use of matching.

Researchers have questioned whether one is to use matching at all (see e.g. Jarrold and Brock, 2004). Covariate analysis is recommended instead of matching when, due to a large difference between group characteristics, matching would distort the groups so much that they would no longer represent their respective groups faithfully (Seltzer et al., 2004; Tager-Flusberg, 2004). For example, Jarrold and Brock (2004) detail how matching subjects with autism from a wide range of intellectual abilities on IQ could easily reduce the autism group to only those with very high IQ scores. They also point out (see p.84) that "matching groups on more than one criterion is

often extremely difficult and, even if possible, will involve such a degree of selectivity that the generalizability of the findings will be reduced considerably." By the latter part of the statement, they may be referring to the difficulty of doing so without excluding many subjects, in which case the makeup of the resulting groups is not characteristic of the original groups. They may also have referred to the fact that matching on multiple variables is non-trivial. On the other hand, when there is very little residual variance and thus the estimates would mostly introduce noise, then matching can be applied, but it probably does not need to remove many subject. One can also use a combination of these techniques: Matching can augment covariate analysis, as it does not assume a particular model between the confounders and other variables (e.g. linear relationship), whereas covariate analysis can take care of the group difference left over after matching (Tager-Flusberg, 2004). Blackford (2009, p.349) argue that matching has advantages, including that "matching produces effect-size estimates with smaller variance than covariate adjustment, analyses on matched data are more robust, and matching can control for more confounders than covariate adjustment, for a given sample size."

Perhaps the most widely known matching approach is *pair-matching*, introduced by Rubin (1973) and popularized especially by Rosenbaum and Rubin (1983). The idea behind the approach is that if we could know the outcome in a person's life both if s/he received a treatment and if s/he did not, then we would know exactly what the effect of the treatment is. Although this is clearly impossible for us to do, yet we can approximate it in the following way: First we choose subject pairs that were very similar on all relevant measures before one of them received the treatment. We can assume that any difference arising between them later is the effect of the treatment. One of the subjects then receives the treatment, and afterwards we measure how the outcome variables changed. If we do this for many subject pairs, we can calculate a quantity known as the *Average Treatment Effect for the Treated subjects* (ATT), which is an estimate of the effect of a treatment on the dependent variable(s). The goal of pair-matching is thus to select controls (who did not get the treatment) for all treated subjects, such that the paired subjects are very similar on the variable(s) used for matching. It can be a 1-to-1, 1-to-N, or M-to-N matching, as one can compare outcome averages when multiple subjects are selected. As a notable side-effect, finding well-matched subject pairs results in a good balance between groups of the selected subjects overall (Gu and Rosenbaum, 1993). The word "treatment" can be understood in a general way to mean many different things, including an actual drug treatment, partaking in training, a change in an eating habit, or even developing a disorder. Thus the above approach can be used not only for experimental but also for observational studies, more specifically case-control studies, of which the CSLU ERPA Corpus is an example. It is even more important for the latter, since, unlike in experimental studies where

one can assign subjects randomly to the treatment group, here one does not have control over treatment assignment. Several implementations for pair-matching are available, including the R packages `Optmatch` (Hansen, 2007), `MatchIt` (Ho et al., 2011), and `Matching` (Sekhon, 2011) for two groups, and `twang` (Ridgeway et al., 2014) for multiple groups.

Pair-matching is generally performed based on *propensity scores*, which are the probability of a subject being assigned to the treatment group, as using them results in an unbiased estimate of the treatment effect. As Dehejia and Wahba (2002, p.151) states: "When the relevant differences between any two units are captured in the observable (pre-treatment) covariates, which occurs when outcomes are independent of assignment to treatment conditional on pre-treatment covariates, matching methods can yield an unbiased estimate of the treatment impact." Blackford (2009) gives other reasons why matching on propensity scores is desirable; for example, more subjects can be kept than when matching on multiple covariates, thus the groups are more similar to the overall population and less bias is introduced. It has another practical advantage, namely that it is much easier to match on one variable than on multiple dimensions (Smith and Todd, 2005). An issue with using propensity scores is, however, that two subjects can have very similar propensity scores even when their characteristics differ widely. In other words, propensity score matching does not guarantee that the paired subjects will be impressionistically similar. For this reason, some researchers prefer combined approaches where one uses the propensity score when the paired subjects have quite similar covariate values according to some predefined condition.

Propensity scores are to be derived from pre-treatment variables (see e.g. Rubin, 1991), especially those that may be relevant for the outcome variables as well. Blackford (2006, p.98) explains it this way: "Only variables that are expected to be related to both group assignment (e.g., sibling with or without Down syndrome) and the outcome variable, but not caused by either, should be included. Gender and birth order are examples of appropriate variables." Propensity scores can be estimated from multiple pre-treatment variables with a logistic regression model.

Unfortunately, we may not always have enough information for estimating propensity scores, and using variables that are themselves affected by the treatment or variables that are not related to the outcome would result in *overmatching* (matching that is superfluous or erroneous, thus harming the statistical efficiency or the validity of the study; see e.g. Rothman et al., 2008). For example, we can consider autism as a "treatment-effect". It is not entirely clear what we could use as "pre-treatment" variables, but as stated above, these are factors that are related to autism risk and cannot themselves be affected by whether the subject is autistic or not, but should be potentially relevant for the dependent variable of interest. Some factors have been shown to be related to autism risk, including gender, the number of affected siblings, birth order, scores of

the parents on cognitive measures, genetic issues such as de-novo mutations, and the presence of certain environmental toxins and air pollution during gestation (see e.g. Chaste and Leboyer, 2012). But we may not have sufficient information about them in the study, which seems to be the case for the CSLU ERPA Corpus as well. When we do not have that information, it may be tempting to use whatever we have to estimate propensity scores, such as cognitive measures. Note however that autism is often comorbid with intellectual disability (in about 50–70% of all cases; see Matson and Shoemaker, 2009). So we cannot exclude the possibility that intellectual ability is affected by having autism (although it could be the other way around as well, or both could be the result of a common underlying neurological condition). Using IQ measures to estimate propensity scores could, therefore, result in overmatching.

When one cannot estimate propensity scores due to a lack of suitable pre-treatment covariates, it still makes sense to use matching on post-treatment ones. Even though that does not guarantee getting an unbiased estimator of the treatment effect, it ensures that any difference found is not due simply to a difference in these variables. For example, for children with a neurodevelopmental disorder, the "post-treatment" covariates can be cognitive and language abilities. Doing our analyses matched on these does not ensure that we measure the actual effect of developing the disorder, but rather that any differences found are not mediated by the level of these abilities (which in turn may have been affected by having the disorder), but are due to the disorder through some other mediators.

Regarding what variables to use, Mervis and Klein-Tasman (2004) bring up valid reasons against using age-equivalent scores, and argue that one should use standardized scores, and adds that one should at least match for age. Moreover, they maintain that child subjects should come from a relatively narrow age range where one can expect a similar level of development: Their abilities hardly ever develop linearly with age, and thus controlling for age in a linear model is not enough.

When one cannot estimate propensity scores, another type of matching may, in fact, be more appropriate than pair-matching: *Group-matching* ensures that the distribution of covariate values is similar between subject groups (instead of being matched on the level of individuals as in pair-matching), by achieving that summary statistics of the distributions are not significantly different at some given α level. The statistics used are usually just the means (this is called mean-matching by Rubin, 1973), but sometimes other properties are used as well, such as the variance of the distribution. Mervis and Klein-Tasman (2004) proposed using $\alpha > 0.5$, also mentioning $\alpha > 0.2$ as a value to consider. Rubin (1973) examined both matching approaches and found that group-matching works well when the dependent variable is linearly related to the matching variables and results in closer matched overall group characteristics. Shaked and Yirmiya (2004, p.37) also found

that "larger effect sizes were yielded when participants with autism were matched on a group basis, rather than on a one-to-one basis, with the comparison participants." Moreover, matched groups can easily be used with standard statistical techniques, such as mixed effect linear models (see Section 3.4.2), and by transforming the covariates to be linearly related to the target variable, one can satisfy the prerequisite pointed out by Rubin (1973).

One issue with group-matching is that it is not trivial when using more than one variable, which is usually the case when propensity scores cannot be calculated. In a 2004 special issue of the Journal of Autism and Developmental Disorders on matching strategies, the technical aspect of matching was mentioned in only one paper to the best of our knowledge: Mervis and Klein-Tasman (2004) described their procedure for matching on one variable, which involves gradually removing subjects with the lowest or highest scores. But it is unclear if they used a computerized algorithm for this; moreover this may be suboptimal as we will see later. We are unaware of any prior work that describes how to perform group-matching on multiple variables.

In this chapter, we deal with the above issue. We introduce and evaluate multiple algorithms for group-matching using complex matching criteria involving several covariates. Kyle Gorman and I have also made available our implementation of those algorithms to the research community in the form of the `ldamatch` R package, which is the only implementation available for this task to the best of our knowledge. We apply those algorithms to our corpora to come up with groups whose distributions are well-matched after losing only a few subjects, which we use in our analyses in later chapters.

5.2 Matching as an Optimization Problem

The problem we need to solve for group-matching is the following: Let us say we have G groups containing a total number of N subjects, with group membership for the subjects indicated by $g_1, \ \ldots, \ g_N$. Each subject also has an associated covariate vector $\overline{c}_i, i = 1 \ \ldots \ N$. We are looking for the optimal subset of subjects selected by the boolean indicator variables $s_1, \ \ldots, \ s_N$ that satisfy our criteria. The criteria comprise a set of statistical tests that typify the similarity of the covariate distribution between the groups, and possibly the expected group size proportions or the maximum number of subjects that can be removed overall or from particular groups. The statistical tests $t_j, j = 1 \ \ldots \ T$ are each a function from g_i and $\overline{c}_i$ for the selected subjects to a p-value in the real interval $[0 \ \ldots \ 1]$, each of which must give $p_j > \alpha_j, j = 1 \ \ldots \ T$. Since we do not take into account the dependent variables being studied, we can generate multiple matched group configurations and choose the one that is the best according to our criteria.

Let us realize that this is a mathematical optimization problem where we are looking for the optimal integer (more specifically boolean) values for variables $s_1, \ \ldots, \ s_N$. If we were to evaluate all possible combinations, it would be 2^N cases, that is, exponential in the number of subjects N. Naturally, we are interested in the solutions with the highest total number of subjects, so finding a solution with n subjects makes it unnecessary to evaluate combinations with counts smaller than n. Evaluating all cases while removing up to n subjects comprises $1 \ + \ N \ + \ \binom{N}{2} \ + \ \ldots \ + \ \binom{N}{n}$, that is $\sum_{i=0}^{n} \binom{N}{i}$ cases, which soon becomes intractible if many subjects need to be removed. In the absence of a mechanism to exclude cases that are guaranteed not to be along an optimal path, excluding a part of the search-space can result in finding a suboptimal solution, as there can be many local minima for this problem type.

This is a discreet optimization problem, more specifically an integer programming problem (but not necessarily an integer linear problem, depending on the constraints), which is NP-hard (Nemhauser and Wolsey, 1988). One can attempt to transform it into a linear programming problem, which can be solved in polynomial time using a *branch and cut* algorithm, by searching for real values instead of integers and then deriving integers from those. But the solution may not be optimal or even feasible, and using diverse statistical tests as criteria would further complicate it. We pursue a different avenue here: First we represent the search-space in a meaningful way and then come up with heuristics that only search in a subset of the enormous search space to find an acceptable solution.

5.3 Matching Algorithms

Below we describe several search strategies that evaluate a subset of all possible subject configurations (specified by $s_i, i = 1 \ \ldots \ N$; see Section 5.2) with the goal of finding one with a non-significant difference between the groups at a given level. For a set of T statistical tests $t_j, j = 1 \ \ldots \ T$, we say that the difference is non-significant if the p-value p_j from test t_j for the groups is above a pre-specified threshold α_j (α_j can be e.g. 0.2 or 0.5). We walk the search space with the aim of optimizing the following measures, in decreasing order of importance: First, we want to keep as many subjects as possible. Second, we either want to maintain the ratio of the group sizes close to a given ratio (such as the original group size ratio), or we prefer to keep subjects in certain groups more than in others by setting up a preference order among the groups. Third, we want to minimize the difference between the groups, by maximizing the minimum p-value–threshold ratio

r that occurs for any test criterion:

$$r = min_{j=1 \ldots T} \frac{p_j}{\alpha_j}.$$

Note that $r \geq 1$ if and only if the groups are matched (see Section 5.2). When comparing possible matched subject configurations, the one with better metrics in this order of priority is preferred. Solutions for which the above are identical are considered equivalent. We have implemented several algorithms in the `ldamatch` R package (see Section B.1.1). Here we describe the algorithms, referring to them by their name in `ldamatch`.

5.3.1 Random search (`random`)

This algorithm was conceived and implemented by Kyle Gorman. It randomly samples the search space for a given number of iterations choosing the subjects to keep randomly according to the binomial distribution, gradually decreasing the expected value of their count from N to G. The search stops after the specified number of iterations I and yields the best solutions found. It is a non-deterministic algorithm with $O(I \cdot T)$ running time, which depends only on the required number of iterations and on how long it takes to evaluate the criteria for any particular subject configuration.

5.3.2 LDA-based heuristic search (`heuristic1`)

This algorithm was first suggested by Jan van Santen and implemented by Kyle Gorman. Its basic idea is to do a dimensionality reduction by projecting the covariates onto one dimension, since it is much simpler to do the matching when there is only a single dimension, as we noted above. While there are many ways to do this dimensionality reduction (including PCA), Fisher's Linear Discriminant Analysis (LDA; Rao, 1948) is somewhat unique in that it allows us to incorporate the dependent variable for the best separation of classes. Specifically, it maximizes the ratio of the interclass variance to the ratio of the within-class variance.

The matching algorithm excludes one subject at a time with the most extreme mapped value, always keeping the proportion of the group sizes around its original value. The search stops when the criteria are satisfied. It is a deterministic algorithm, with a computational complexity of $O(N \cdot T)$ (linear), where N is the total number of subjects, as it evaluates only a tiny, but promising part of the search space.

In our implementation, we made it possible to specify some additional options for the search: The algorithm can take into account r when deciding which subject to remove next of the two at

either end of the available range, and removes the one for the higher r. It can also consider for removal all subjects, or exclude a center portion of the subjects from consideration.

5.3.3 Test-statistic based heuristic search (`heuristic2`)

This algorithm was conceived by Jan van Santen and is a constructive algorithm (it constructs a solution in a series of steps, always taking the step that seems to take it nearer to a solution; see for example Genova and Guliashki, 2011). The basic idea is that we use the value of r to decide which way to proceed when walking the search space, that is, which subject to remove next to attain the largest improvement in the target criteria. In every step, it calculates the r value that results from removing each remaining subject in turn and then removes the one with the highest r. If multiple configurations with the same number of subjects meet our criteria, we take advantage of the other metrics to rank them and choose the first one. It is a deterministic algorithm, with a computational complexity of $O(N^2 \cdot T)$ (quadratic in the number of subjects).

Intuitively an issue with this algorithm is that it is not able to proceed in the right direction toward the global optimum when that results in a local drop in r. For example, when it needs to remove two subjects with extreme covariate values on the opposite ends of the scale, the removal of either subject makes the group balance worse, whereas the removal of both subjects at the same time may improve it. We addressed this issue by introducing the `heuristic3` and `heuristic4` algorithms.

5.3.4 Test-statistic based heuristic search with look-ahead (`heuristic3` and `4`)

These algorithms are an extension of the above `heuristic2` algorithm by the author of this dissertation to look ahead several steps. Look-aheads have been utilized for various problems, including vehicle routing problems (Atkinson, 1994), decision-tree induction (Dong and Kothari, 2001), and finite-state transducer compositition (Allauzen et al., 2010), but we are not aware of their approach having been applied to the matching problem. The difference between `heuristic3` and `heuristic4` is how they decide on which subject to remove next from among the possible candidates. What is common between them is how they come up with these candidates.

The procedure first identifies one or more sets of L subjects, denoted here by S, whose removal results in the biggest improvement L steps down the road (more than one set if they are equivalent on our metrics), then it removes one subject from those sets. Note that it is possible to reach one of the best sets L steps down the road, but it does not commit to removing L particular

subjects at this point, as it may find a still better combination as it progresses. Then it repeats the process starting with one less subject, so looking one step further. Which subject to remove next is decided differently by the two algorithms: `heuristic3` decides solely based on the r values, while `heuristic4` prefers to eliminate a subject that is a candidate for removal in the highest number of subject sets (in a sense, making as small a commitment as possible, as it removes a subject that is likely to be removed at some point during the process).

The algorithm for choosing the next subject for removal in `heuristic3`:

1. Let l be L and C be S.
2. If $l = 1$, choose the subject to be removed randomly from C and exit.
3. Let l be $l - 1$.
4. Let S_l be all n subject subsets of size l from C.
5. For each subject subset in S_l, calculate r with its subjects removed.
6. Let C be the subject sets from S_l with the highest r value.
7. Go to Step 2.

The algorithm for choosing the next subject for removal in `heuristic4`:

1. Count the number of times each subject occurs in S and keep the ones with the largest count as candidates.
2. If there are more than one candidate, calculate r for each one and keep the ones with the highest r.
3. If there are more than one candidate, choose one of them randomly.

The algorithms are non-deterministic (as they choose randomly among equivalent options), they follow a depth-first search strategy, and their complexity is $O(N^{L+1} \cdot T)$ where N is the number of subjects and $L \geq 1$ is the number of steps the algorithms look ahead. For $L = 1$, they are both equivalent to `heuristic2`.

In our implementation, we make it possible to remove multiple subjects in each step (until the remaining number of subjects decreases to a certain number), so as to make it feasible to work with groups containing thousands of items. The reason is that it can be time-consuming to calculate the r value for all items, and these values may not change much after removing one item. So when we work with many items, calculating r only once in S steps may not degrade the quality of the matching much, but speeds up the search by a factor of S. This of course does not change the

assymptotic complexity of the algorithm, yet can be important in practice, especially with these algorithms that have supra-linear complexity (e.g. using $S = 100$ may reduce the running time from 100 hours to about an hour).

5.3.5 Exhaustive search (`exhaustive`)

Doing an exhaustive search seems to be the simplest possible approach, but it is computationally prohibitive for large search spaces, as mentioned earlier. We implemented an exhaustive search algorithm that makes the search feasible when only a few subjects need to be removed to reach well-matched subject groups. We can estimate an upper bound on the number of subjects that need to be removed using one of the heuristic algorithms, and based on that we can decide if it is feasible to perform an exhaustive search. Having this algorithm at our disposal not only can give us the optimal solution, but it also enables us to assess how well other approaches fare.

Let us look at an example. Given two groups, each containing 20 subjects, and assuming that the computer can process 1,000 out of the 2^{20+20} subject configurations per second (which is over $1.099 \cdot 10^{12}$ configurations), evaluating all cases would take over 34 years. However, if a heuristic algorithm finds a solution that meets the matching criteria by removing five subjects, then we know that the search will complete in 13 minutes or less, which makes running an exhaustive search feasible. If it turns out during the process that the optimal solution requires the removal of only three subjects, then it will finish in less than 11 seconds.

We implemented the exhaustive search algorithm as a breadth-first search: It explores the cases in the order of decreasing goodness. Our criteria rank candidate solutions first based on the total number of subjects retained. Ties are broken in favor of smaller divergence from the desired group proportions measured using the Kullback-Leibler divergence (K–L; see e.g. Cover and Thomas, 2006), and finally by the value of r. When multiple solutions are available with the same size and K–L divergence, we favor those with higher values of r.

5.4 Matching the CSLU ERPA Corpus Subjects

In this section, we describe the application of the matching algorithms to the subject pool of the CSLU ERPA Corpus from a technical viewpoint. This also enables us to compare the algorithms on real-life data. First we summarize our goals for the matching task, then we describe a process for finding the best possible solution within the limits set by the available computational resources. Finally we present some properties of the results and compare the algorithms.

Here we do not go into detail about the significance of this problem or even the abbreviations we

use, as we have done that elsewhere and one does not need to understand those to understand the technical aspect. For a detailed description of the groups, the covariates, the matching criteria, our motivation, and the rational for matching the subjects this way, see Section 4.2, especially subsection 4.2.4.

5.4.1 Matching criteria

Our goal was to find four sets of subjects from four groups (TD, ALN, ALI, SLI) such that different group pairs are matched on different covariates:

- all groups on age
- SLI and ALI on PIQ and VIQ
- ALI and ALN on the ADOS score
- ALN and TD on PIQ and VIQ

We wanted to keep all subjects with SLI, and as many as possible for the ALI, ALN, and TD groups, in decreasing order of preference. The p-value for multiple test-statistics needed to be at or above the significance level $\alpha = 0.2$ (two-tailed). Specifically, we wanted to match the group means using a t-test with unequal variances, and the overall shape of the distributions using the Anderson–Darling test, as for example Facon et al. (2011) show that it is advisable to match the overall shape of the distributions as well.

If we match each group pair independently of the other pairs, generally we would find different subsets of subjects for them, which is not an acceptable solution for us: We wanted to find just one set of subjects that meets all of the criteria above at the same time, while also optimizing the group size and other metrics (see Section 5.2).

5.4.2 Approaches to matching subsets of groups on different variables

We tackled the problem of finding one set of subjects that meets all of our criteria at the same time two different ways: First, we matched one group-pair at a time, using the previously matched groups in subsequent matchings (see more below); we call this *consecutive matching*. It turned out that it would require an enormous effort to find quality solutions — most importantly ones that preserve a substantial number of the subjects — using this approach. Second, we matched the groups simultaneously, globally optimizing the solution. This yielded reasonable results with a much smaller investment, other than having had to create a complex matching criterion function. In the following, we review both approaches and their results.

Consecutive matching

As stated above, the basic idea was to match pairs of groups at a time, building on the result of the previous matches:

Step 1: Match SLI and ALI, preserving all available SLI subjects, as this is the group with the smallest number of available subjects in the CSLU ERPA corpus.

Step 2: Match ALI and ALN, using the ALI group from Step 1 unchanged, only removing subjects from the ALN group.

Step 3: Match ALN and TD, using the ALN group from Step 2 unchanged, only removing subjects from the TD group.

This approach turned out to be infeasible, for the following reason. When we perform these steps with a deterministic algorithm (e.g. `heuristic1` or `heuristic2`), the solution may be particularly low quality. When we perform these with a non-deterministic algorithm (e.g. `heuristic3` and `heuristic4`) or one that provides all equivalent solutions (e.g. `exhaustive`), we get multiple solutions in each step, which gives us an opportunity to optimize the final result. One way is to randomly choose one of the solutions as the basis for the next step; this however generally resulted in rather small overall group sizes. Another way is to evaluate the best ones from each step; but this turned out to be infeasible in a limited amount of time. Moreover, even if we do that, it may not even approximate the globally optimal solution: One may need to choose a suboptimal solution in an earlier step to find a larger number of total subjects down the way. For example, removing one more subject in Step 1 may result in being able to keep many more subjects in Steps 2 and 3. For these reasons, after investing a substantial amount of time, we abbandoned the consecutive matching approach. Based on our experience, it seemed obvious that we needed to employ a global optimization strategy: one that takes all our criteria into account simultaneously.

Simultaneous optimization

The basic idea of this approach is rather simple: We create one complex set of matching criteria that contains everything we require in the final solution, and then use that to work with all subjects from all groups at the same time. We just needed to create the necessary infrastructure components for attaining this, which includes the ability to:

1. specify complex sets of matching criteria, including different criterion functions and thresholds for different subsets of the groups;

2. calculate r (the minimum p-value–threshold ratio) for the complex matching criteria;

3. calculate evaluation metrics for how well a set of subjects suits the matching criteria;

4. set up an ordering among the possible subject configurations; that is, be able to decide which one of two subject sets has better evaluation metrics.

5.4.3 The matching process

Since every matching algorithm has its own respective drawbacks and benefits, we do not commit to using any one of them, but use each one if possible. We use the exhaustive search only when it seems feasible based on the best result from the other algorithms. Finally, we keep the matched tables with the best evaluation metrics (see Section 5.2).

It can be important from a practical viewpoint that the heuristic algorithms may find better solutions for stricter criteria. For example, it may be able to preserve more subjects when matching the shape of the distributions and not only their means, or when required to keep all subjects from one of the groups. This may happen because they only consider a small part of the search-space and the stricter criteria can help to guide them into a direction that will prove to be better long-term. In other words, additional criteria may help to make better local decisions and thus to attain better global decisions. We can take advantage of this and use the best of the candidates from the outputs for a large set of matching criteria.

5.4.4 Results

We applied the simultaneous optimization approach and the process outlined above to find the best subject configurations that meet our criteria for the CSLU ERPA ADOS Corpus. From a researcher's point of view, the important thing is that the studies using the matched groups be well-powered. Here we are also interested in what the advantages and drawbacks of each algorithm are.

The results are given in Table 5.1. We put in bold the best values for each aspect of the solutions found. The "hA la. L" abbreviation in the column headers (e.g. "h3 la. 2") refers to the heuristic algorithm number A with look-ahead L. The p-values are the smallest ones from all solutions found. Depending on the chosen criterion, the divergence value can either be the Kullback–Leibler divergence (see e.g. Cover and Thomas, 2006) of the group sizes from the expected proportions, or a number that is larger if more subjects are removed from groups that we prefer to keep unchanged. For the problem at hand, it was the latter, as for example we preferred to keep as many subjects as possible in the SLI group; a lower divergence means that subjects were removed from groups with larger counts. Instead of going into details about how we calculate the divergence value, we

normalized them in the table by dividing each one by the largest divergence in the same row. We did not include the results of the `heuristic1` algorithm as it was originally designed for matching two groups and it always failed for this complex set of criteria. The `random` algorithm practically always finds a solution, but the quality of the solution is not comparable to that of the heuristic algorithms. The `exhaustive` search was not feasible for even the smallest of these problems.

We ran the algorithms on a cluster of computers parallelly. The `heuristic3` and `heuristic4` algorithms ran in multithreaded mode. The machines were commodity x86–64 machines running Ubuntu Linux. The number of logical cores for running threads was between 16 and 24, and the CPU MHz ranged from 1600 to around 3000.

Table 5.1: Information on matching results when $p >= 0.2$ for both the t-test and the Anderson–Darling test. The fields contain: number of excluded subjects, divergence from the optimal balance compared to the worst one, number of solutions returned, minimum p-value among all solutions after matching, search time in hours (and compared to the worst one).

	`h2`	`h3` la. 2	`h3` la. 3	`h4` la. 2	`h4` la. 3
CA	**8 excluded**; div.: 1.00; 1 solutions; p=0.23; 0.0 hours; (0%)	**8 excluded**; div.: 1.00; 70 solutions; p=0.22; 0.6 hours; (3%)	**8 excluded**; div.: 1.00; **123 solutions**; p=0.21; 21.8 hours; (100%)	**8 excluded**; **div.: 0.80**; 34 solutions; p=0.22; 0.2 hours; (1%)	**8 excluded**; **div.: 0.80**; 64 solutions; p=0.21; 7.2 hours; (33%)
CA–VIQ	15 excluded; div.: 0.50; 1 solutions; p=0.21; 0.1 hours; (0%)	**14 excluded**; div.: 1.00; 5 solutions; p=0.20; 0.4 hours; (2%)	**14 excluded**; div.: 1.00; **6 solutions**; p=0.20; 21.9 hours; (100%)	**14 excluded**; div.: 0.75; 1 solutions; p=0.21; 0.6 hours; (3%)	**14 excluded**; **div.: 0.25**; 2 solutions; p=0.20; 13.4 hours; (61%)

	`h2`	`h3` la. 2	`h3` la. 3	`h4` la. 2	`h4` la. 3
CA–	15 excluded;	**14 excluded**;	**14 excluded**;	**14 excluded**;	**14 excluded**;
VIQ–	div.: 0.50;	div.: 1.00;	div.: 1.00;	div.: 0.75;	**div.: 0.25**;
PIQ	1 solutions;	5 solutions;	**6 solutions**;	1 solutions;	2 solutions;
	p=0.21;	p=0.20;	p=0.20;	p=0.21;	p=0.20;
	0.1 hours;	4.0 hours;	17.5 hours;	2.3 hours;	188.1 hours;
	(0%)	(2%)	(9%)	(1%)	(100%)
CA–	18 excluded;	**17 excluded**;	**17 excluded**;	18 excluded;	**17 excluded**;
VIQ–	div.: 0.81;	div.: 1.00;	div.: 0.60;	div.: 0.60;	**div.: 0.60**;
ADOS	1 solutions;	18 solutions;	48 solutions;	**86 solutions**;	2 solutions;
	p=0.21;	p=0.21;	p=0.20;	p=0.21;	p=0.20;
	0.1 hours;	5.6 hours;	185.0 hours;	2.5 hours;	192.0 hours;
	(0%)	(3%)	(96%)	(1%)	(100%)
CA–	18 excluded;	18 excluded;	**17 excluded**;	18 excluded;	**17 excluded**;
VIQ–	div.: 1.00;	div.: 0.99;	div.: 0.74;	**div.: 0.74**;	**div.: 0.74**;
PIQ–	1 solutions;	**57 solutions**;	48 solutions;	8 solutions;	2 solutions;
ADOS	p=0.21;	p=0.21;	p=0.20;	p=0.21;	p=0.20;
	0.2 hours;	8.3 hours;	22.7 hours;	0.5 hours;	25.0 hours;
	(1%)	(33%)	(91%)	(2%)	(100%)

5.5 Discussion of the Matching Algorithms

Based on the application of the solution to the CSLU ERPA Corpus (see Table 5.1), we can see that there is not a large difference between the algorithms in the number of subjects retained: Of the 113 candidate subjects, the best solutions preserved at most one more. What still differentiated solutions was how much the size of the groups diverged from what was expected. The `heuristic3` and `heuristic4` algorithms often found multiple equivalent solutions. This can be useful because the researcher can check if his or her findings stand up for various subject configurations that are matched at almost exactly the same level (only the p-values may vary a bit, but they all are above the threshold). Note that there is a random element to the solutions, as described earlier, so

running them again may result in slightly different outcomes. For example, running the algorithms with a look-ahead of 2 multiple times increases the chance of finding the optimal solution.

The largest difference between the algorithms was unfortunately in their running times: The simplest `heuristic2` algorithm finished in a matter of seconds or minutes, while the algorithms with larger look-aheads ran for hours or days on multi-core machines using all cores. (The durations are suitable for comparing their order of magnitude only because we ran the algorithms on a cluster of machines with varying processing speeds.) Nevertheless we believe that it may be worth running such algorithms for a few days, or even longer, even if we can preserve only one or two more subjects, as each subject comes with a substantial marginal cost, and the result may get used in multiple publications in a course of several years. Of course, for other data sets the difference in the solutions may be larger or smaller than what was observed here.

The main value of the approach to matching presented here is that it is a systematic approach that has been made available publicly for use by the research community. The matching algorithms are part of the `ldamatch` packages, which has been available on `CRAN` (the central R repository) since early 2016. We plan on publishing the code that implements our approach for finding solutions for complex matching criteria in the future.

Possible research directions include analyzing the matching problem as a linear programming problem and then finding a boolean solution around the point specified by the real values. For improving the current solutions, a local-improvement algorithm (Aarts and Lenstra, 2003) can be used, for example after finding a solution with the one of the heuristic algorithms.

5.6 Summary of our Work on Matching

We made a case for matching subject groups on overall properties, such as cognitive measurements of the subjects, as the only feasible approach in certain situations. We designed and implemented a systematic approach for matching multiple groups using complex criteria. Evaluation of the algorithms showed that even the fastest ones can find acceptable solutions, which can be improved upon by more computationally intensive approaches. We made the algorithms available to the research community. The result of matching the subjects of the CSLU ERPA Corpora forms one of the bases for this work, as described in Section 4.2.4.

Chapter 6

Acoustic Characterization of Prosody

The earlier chapters provided the theoretical foundation and the data for our analyses, including the speech corpora with subject groups from several diagnostic groups matched on relevant cognitive measures. Now we are ready to start analyzing this data, first by calculating diverse acoustic-prosodic features. We compare these directly between the groups to identify significant difference, and also deduce qualitative differences in the shape of the intonation curves.

6.1 Motivation for the Acoustic Analysis of Speech Prosody

Even when the speech of children with ASD is functional, its prosody can still be different from that seen in typical development. For example, van Santen et al. (2010) found that even though autistic subjects in their subject pool were able to produce functionally correct distinctions (e.g. between sentence types, or the words with focus stress), yet they had quantitative differences in the way they expressed these distinctions compared to typical development. More specifically, when expressing stress, the children in the ASD groups used durational cues disproportionately less than those in the TD group (see also Section 2.3.6). Such findings imply that there must be high-level features of prosody, including summary statistics, that differentiate the speech of children with ASD from that of their typically developing peers.

Our purpose in this chapter is to try to identify acoustic features that help in characterizing speech prosody in ASD. Such features can also be used for classification and regression to differentiate between diagnostic categories. If we can map these features to subjective qualities of prosody, they can help to see how the prosody of autistic children differs qualitatively from typical speech. This knowledge may in turn help in remediation. We are also interested in the relationship of prosody to content. We work with a corpus that has been transcribed and annotated with supplementary information on language use. This knowledge can help to see if prosodic differences can be partly explained by differences in content, or if these two are more-or-less independent of each other.

The approach we used was to analyze the acoustic equivalents of pitch, amplitude, rhythm and pausing, namely the fundamental frequency, intensity, and segmental durations using multiple computational approaches. The computational methods captured different aspects of prosody, such as the statistical properties of F0 and intensity and the shapes of the utterance intonation curves. First, we describe these acoustic phenomena at a high level, then the features we calculated from them. We perform statistical tests with appropriate controls against false discovery to answer the question on whether a given set of features distinguishes ASD from TD or SLI, and what that difference means. We also analyze acoustic features in relation to content features. This chapter

is partly based on work we published earlier (Kiss et al., 2012; Kiss and van Santen, 2013).

6.2 Acoustic-Prosodic Feature Sets

In this section, we describe how we calculated acoustic features from a corpus of spontaneous speech conversations, the CSLU ERPA ADOS Corpus (see Section 4.2.1). It can be considered to be composed of units of different sizes. Going from the shortest one and increasing their length they are: the speech frames, then the phonetic segments (speech sounds and pauses), syllables, words, C-units (see Section 4.2.5), the utterance, and finally all speech for a subject. We worked at all these levels depending on the purpose at hand, mostly from time series of F0 and intensity derived on a per-frame basis. We covered F0 detection in some detail in Section 3.1. For the intensity curves, we used the logarithm of Root Mean Squared (RMS) frame values. The log transformation gave the intensity features more predictive power when used in a machine learning framework in another experiment (see Chapter 8.2), which suggested that we use it this way. The feature sets were statistical features of prosody, functional features, and parameters of an intonation model.

6.2.1 Statistical features of prosody

Statistical feature set

We defined a set of features that we used regularly for utterances or speakers to characterize the distribution of the values on various measures (such as F0, intensity, or duration). We used this not just in this chapter, but also elsewhere in this work. The statistics were robust and non-robust versions of the first four standardized moments; see Table 6.1. There is no universally accepted standard robust version for coefficient of variation (CoV), skewness, and kurtosis. We chose one of the existing definitions as follows:

- robust CoV: $\frac{\text{IQR}}{\text{median}}$ and $\frac{\text{MAD}}{\text{median}}$
- robust skewness: SK_2 in Kim and White (2004)
- robust excess kurtosis: KR_2 in Kim and White (2004)

Table 6.1: The statistical feature set calculated for prosody curves

statistics	non-robust	robust
location	mean	median
	minimum	10th percentile
	maximum	90th percentile
spread	standard deviation (SD)	inter-quartile range (IQR)
	coefficient of variation (CoV)	robust CoV
		median absolute deviation from the median (MAD)
assymetry	skewness	robust skewness
peakedness	logarithm of excess kurtosis	logarithm of robust excess kurtosis

The created three sets of features: the first one for every utterance, the last two for every speaker:

1) `US`: Per-utterance statistics of the values for each utterance.

2) `SUS`: Statistics of the per-utterance statistics, one set per speaker; since the 3rd and 4th moments require more samples to work with than the lower order moments for reliable estimates, we did not use those here, as there may be too few utterances for their estimation.

3) `GS`: Per-speaker statistics of the values from all utterances together (global statistics). These are the statistics for the concatenated intonation curves of all utterances of a speaker, calculated the same way as in the `US` case.

We calculated these statistics for per-frame information (namely F0 and intensity), as well as phoneme, pause, and syllable durations and counts.

F0 features

We analyzed statistical properties of the F0 information for each speaker in two ways: first, through the feature sets described in Section 6.2.1, and second, by replicating the features that Sharda et al. (2010) and Bonneh et al. (2011) showed to be significantly different between their ASD and TD groups. Sharda et al. (2010) identified the following features that distinguished the groups: mean F0, mean F0 range, and mean "pitch excursion", calculated as

$$\frac{12 \cdot \log_2 \frac{\mathrm{F0_{max}}}{\mathrm{F0_{min}}}}{\mathrm{duration}}$$

(essentially the F0 range in semitones divided by the utterance duration). The F0 features that Bonneh et al. (2011) identified were the F0 range, F0 SD, and the height of the normalized F0 histogram, the latter being their most discriminative feature.

Duration features

We derived segmental durations from forced-aligned transcripts of the corpus created by colleagues Kyle Gorman and Katina Papadakis at CSLU. They excluded the more problematic C-units, namely those that have non-speech sounds, overlaps with the speech of another speaker, or incomprehensible words. For the others, they determined the start and end times of the phonemes and words. They created phonetic transcripts automatically for the manual textual transcripts of the CSLU ERPA ADOS Corpus using the Carnegie Mellon University (CMU) Pronouncing Dictionary and the Sequitur grapheme-to-phoneme system (Bisani and Ney, 2008) for out-of-vocabulary words, and letting the acoustic model resolve homographs. They aligned the phonetic transcript to the waveform recordings using the Prosodylab-aligner forced alignment system (Gorman et al., 2011).

We created a simple rule-based syllabifier (see Section B.1.5) that determines the syllable boundaries for a phoneme sequence. Evaluating its behavior on the alphabetic entries of an automatically syllabified version of the CMU Pronouncing Dictionary (Bartlett et al., 2009) that contains 105,901 syllabified words gave a recall of about 99.8%: it syllabifies all but 196 words in the expected way. Its advantage compared to using a pronunciation list directly is that it can syllabify out-of-vocabulary words as well.

We calculated durational features for each utterance, for all levels: the utterance itself, syllables, vowels, phonemes, and internal pauses. We derived the speaking rate and the articulation rate as the number of syllables divided by the total duration. For speaking rate, the total duration includes sentence-internal pauses; for articulation rate, it comprises the length of the phonemic segments only.

6.2.2 Speaker-specific intonation model parameters

Modeling speech intonation for a specific speaker is important for various applications, including speech synthesis for mimicking the specific speaker (Klabbers et al., 2010), speaker identification or verification (Sönmez et al., 1998), and even for diagnostic purposes, by detecting atypicality in speech prosody (Kiss et al., 2012). Researchers have proposed several intonation models (for a brief review, see Section 3.2). For each intonation model, one needs a method for estimating the model parameters from natural speech of a person. This task is easier for more restricted models,

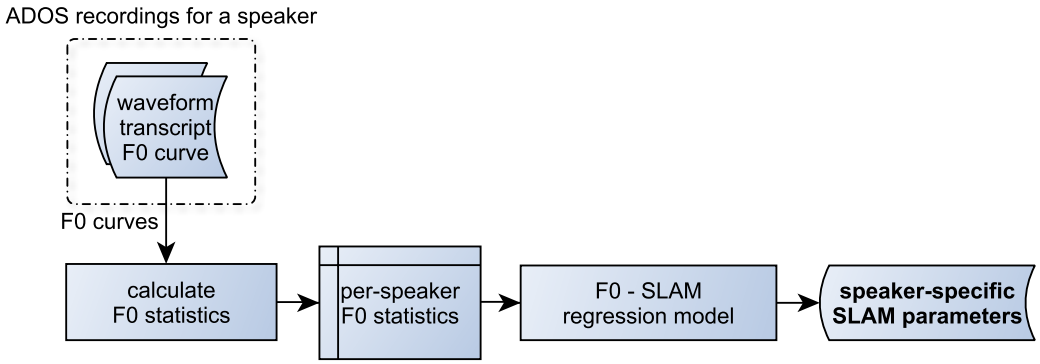


Figure 6.1: Estimating speaker-specific intonation model parameters

such as the Fujisaki model (see Section 3.2.1). For more general models, such as Generalized Linear Alignment Model (see Section 3.2.2), and even for the Specific Linear Alignment Model (SLAM; see Section 3.2.3) the problem is harder. The method developed by Mishra and colleagues for calculating the parameters of the SLAM model from natural speech (Mishra, 2008) requires that the foot structure and the phonetic content be labeled, which requires time-consuming manual work. Making it automatic is desirable, as it can give researchers a tool for deriving the easily interpretable characterization of a speaker's intonation, as well as the possibility to mimic the speaker in a Text-to-Speech Synthesizer (TTS) system.

We estimated speaker-specific intonation model parameters, (i.e. one parameter set for each speaker) in an indirect way from spontaneous speech samples of the subjects in the CSLU ERPA Autism Corpus (see Section 4.2.1). To that end, we created a regression model for estimating parameters of the SLAM model from statistical features of synthetic F0 curves. We used this model to estimate the model parameters (hereafter denoted as "SLAM parameters") for each subject in the CSLU ERPA ADOS corpus, assuming that for a given speaker, the intonation model parameters are constant across the utterances. Finally, we checked to see if these intonation model parameters differ systematically between the groups. See Figure 6.1 for an overview of the process. Below we describe the steps in more detail.

Regression model for estimating SLAM parameters

Approach We generated SLAM parameter sets for "artificial speakers", synthesized the intonation curves for utterances for each speaker, and then trained a regression model to estimate the SLAM parameters from properties of the intonation curves. We chose SLAM parameters and

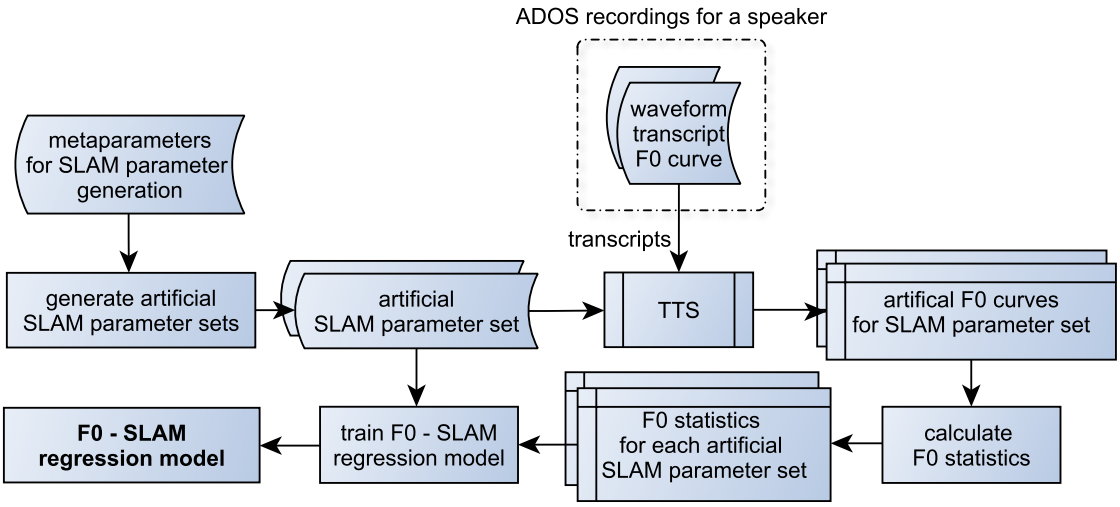


Figure 6.2: Creating regression model for estimating intonation model parameters

utterances that are realistic for the children whose prosody we want to assess. By this we mean that the intonation model parameters are in the range the are likely to include the actual values for our subjects, and the utterance contents are from transcripts of children's speech. Below we describe the steps of this process, depicted in Figure 6.2.

Artificial SLAM parameter sets We generated 2000 sets of parameters for the SLAM intonation model characterizing the prosody of "artificial speakers" (i.e., various speaking styles), as training data for the regression model. We included six of the parameters of the SLAM model in the sets: the three phrase curve and the three accent height parameters (see Figure 3.1). We generated sets of these six parameters randomly based on meta-parameters (or hyper-parameters) that define the range for these. We set the meta-parameters to ranges that are sufficiently wide and which include the range for the young children in our corpus since we want the resulting regression model to work for them. In the speech of neurotypical adults, it generally holds that phrase start $>$ phrase middle (i.e., inflection point) $>$ phrase end; also accent start $>$ accent end $>$ accent middle. However, these constraints may not hold for children, and especially for those with a neurological disorder, therefore we did not enforce such constraints on the generated SLAM parameters. For each phrase parameter, we used the 160–600 Hz range, and for each accent height parameter, we used the 10–260 Hz range.

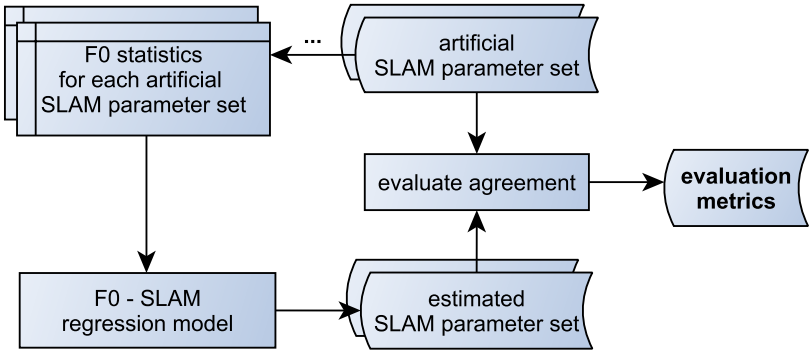


Figure 6.3: Evaluating the regression model for estimating intonation model parameters

Synthesis of Training Data We used the CSLU TTS to synthesize 1000 utterances for each artificial SLAM parameter set. The synthesizer predicts the prosody for synthesized speech using an implementation of the SLAM model and an elaborate duration model (van Santen, 1994). As text to be synthesized, we chose statements randomly from the transcriptions of the CSLU ERPA ADOS Corpus, such that the utterances were different for each SLAM parameter set, to imitate the real situation of having a different set of utterances for each subject in a spontaneous speech corpus. Instead of the more resource-intensive and error-prone process of synthesizing the waveforms and detecting the F0 curves in a separate step, we just synthesized the F0 curves and deleted the parts for unvoiced consonants and pauses. This resulted in 2 million artificial F0 curves as our training data. For the F0 values, we computed the two per-speaker feature sets described in Section 6.2.1.

Evaluating regression models for estimating SLAM parameters

We trained multiple regression models for estimating speaker-specific SLAM parameters with varying amounts of training data and features, then chose the best model. The evaluation of the models took place in a ten-fold cross-validation scheme, using the synthetic F0 curves described above. A more realistic evaluation would be based on natural speech samples that have been labeled with the SLAM parameters, but currently we do not have such a corpus and do not know of one. See Figure 6.3 for an overview of the process.

We experimented with two kinds of regression models:

1. Linear Model with L1 regularization (L1LM)

2. Support Vector Regression (SVR) with a Gaussian kernel

The use of L1LM is attractive in that its coefficients are easily interpretable, it generates sparse models by implicit feature selection during training, and it may be able to extrapolate to output ranges unseen in the training set. It also enables us to create a trade-off between the number of features and the model performance, and thus helps to measure the performance as a function of the number of features. On the other hand, the SVR model can give good results even if the relationship between the predictors and the dependent variable is highly non-linear as long as our features warrant establishing a binary input–output relationship. Unless indicated otherwise, we report the results for the L1LM model because of its advantages, as its performance turned out to be on a par with that of the SVR for estimating the per-speaker SLAM parameters.

Comparison of feature sets We trained models on several F0 feature sets (see Section 6.2.1) and chose the best feature set. As the evaluation metric, we used the Root Mean Squared Error relative to the parameter ranges (RMSE%), averaged over the cross-validation folds. We can see the performances in Table 6.2.

As we can see, the estimation of the phrase curves is reasonable for a wide range of speaker characteristics, but the accent estimation needs to be improved to be useful. The performance is generally better for sets with more features, or at least not significantly worse. See also Figure 6.4 for a visualization of the relationship between the actual and estimated SLAM parameters.

Table 6.2: Performance of several feature sets in RMSE% for estimating SLAM intonation model parameters when trained on 2000 random SLAM parameter sets. `LS` denotes utterance length statistics, the other abbreviations come from Section 6.2.1.

feature set	phrase start	inflection point	phrase end	accent start	accent middle	accent end	mean
F0 `GS`	24.7	23.6	22.3	26.5	28.8	25.6	25.2
F0 `SUS`	12.7	6.5	10.7	21.2	25.5	19.9	16.1
F0 `GS`+`SUS`	10.0	6.2	8.0	19.7	22.4	19.1	14.2
F0 `GS`+`SUS`+`LS`	10.0	6.2	8.0	19.8	22.4	19.1	14.2

Required number of features We trained several L1LM models to examine the performance as a function of the number of features. We varied the number of features by increasing the regularization parameter of the L1LM model gradually from 0 to 1. As we can see in Figure 6.5, the performance seems to plateau at around 40 features for this feature set, but using all the

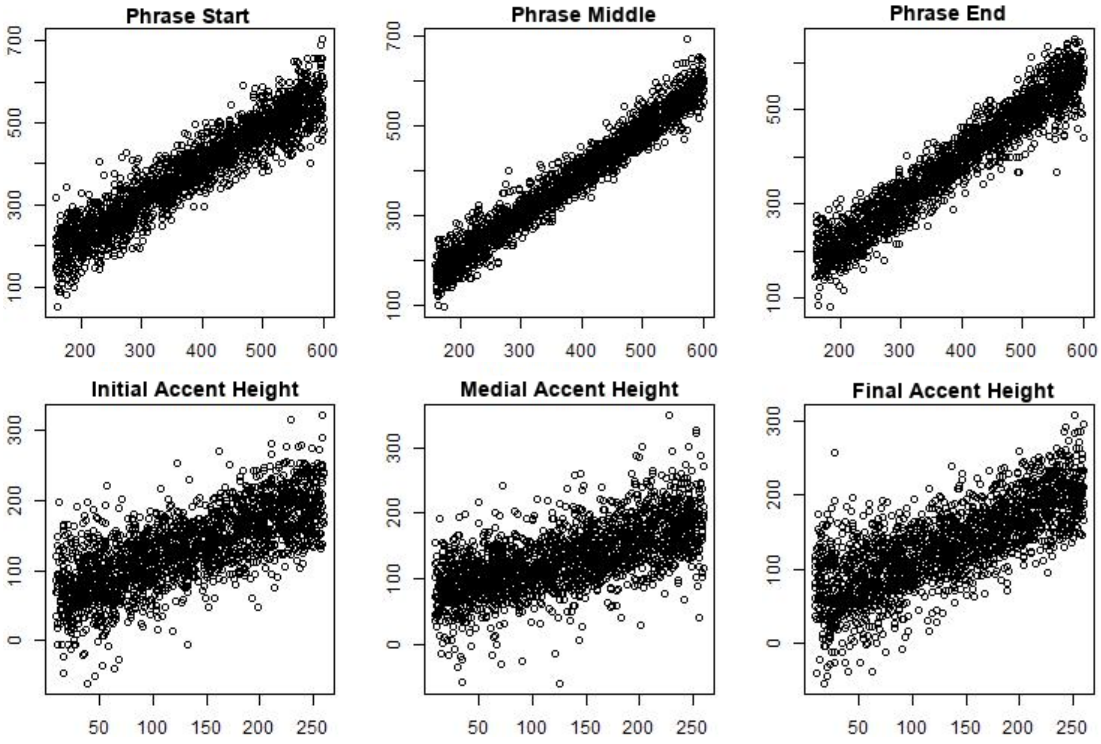


Figure 6.4: The relationship between the actual and estimated SLAM parameters. The actual parameters are on the abscissa, the estimated values are on the ordinate.

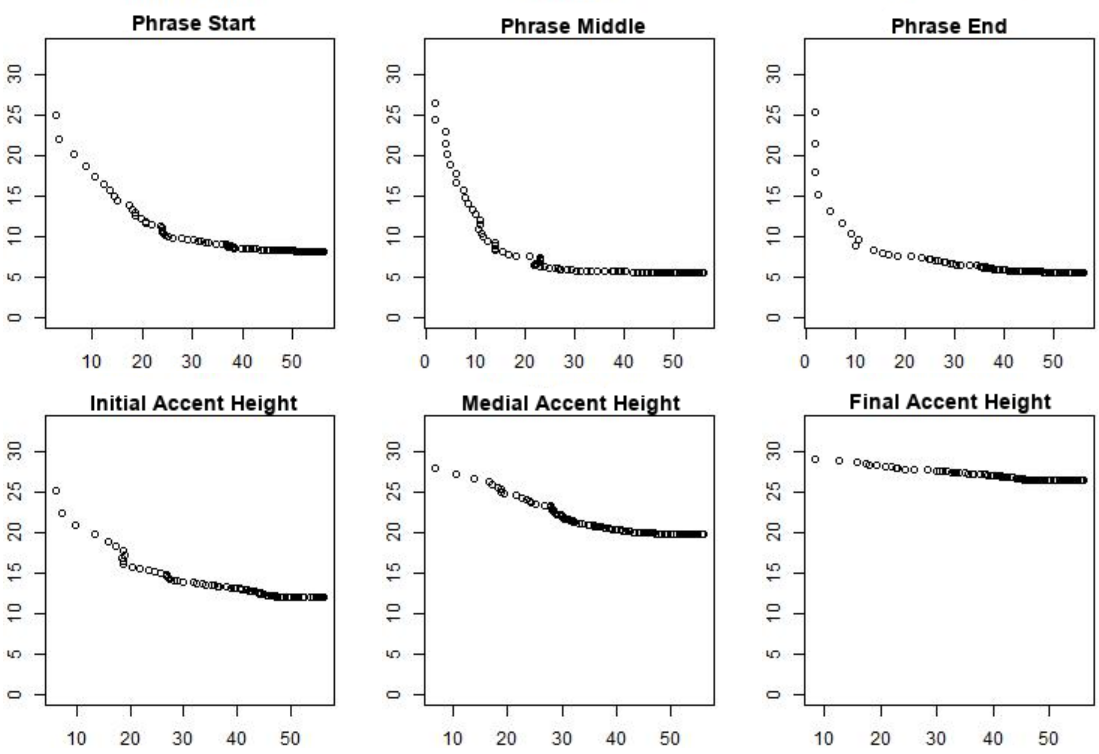


Figure 6.5: SLAM parameter estimation: the performance as a function of the number of features selected by an L1LM model. We varied the number of features by increasing the regularization parameter of the L1LM model gradually from 0 to 1. The number of features is on the abscissa, the performance in is on the ordinate. The performance is measured as RMSE%.

features still increases the performance somewhat.

Required number of utterances We trained several models to examine the performance as a function of the size of the training data. We trained the model on an increasing number of SLAM parameter sets from our training set; see Figure 6.6. The amount of training data we used (2000 artificial speakers) seems to be ample, as there is very small improvement in the performance after about 500 sets.

Potential ways to improve the model Adding more features may improve the performance. For improving the accent estimates, including the F0 peak heights and locations in the feature set may help. Preliminary analyses showed that the per-utterance estimates are very noisy. Using an SVR model that is able to handle the huge amount of synthetic data involved can improve those estimates substantially.

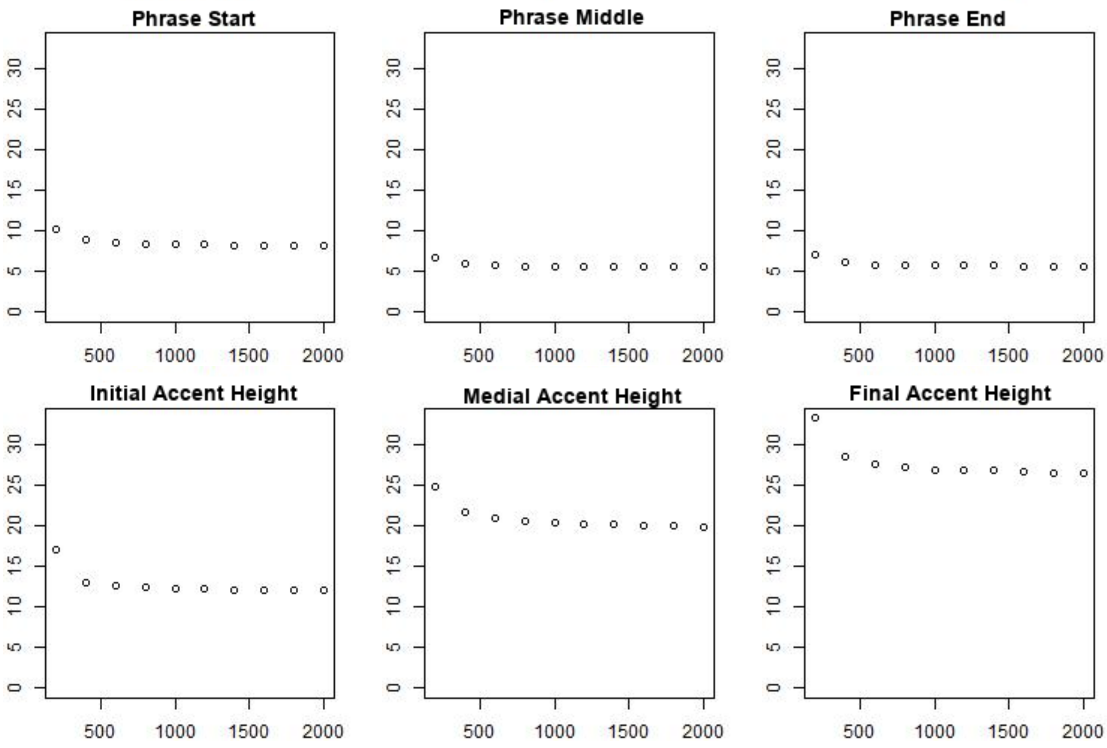


Figure 6.6: SLAM parameter estimation: the performance as a function of the size of the training set in SLAM parameter sets. We trained models using an increasing number of SLAM parameter sets from our training set. The size of the training set (number of artificial speakers) is on the abscissa, the performance in is on the ordinate. The performance is measured as RMSE%.

Estimating SLAM parameters from spontaneous speech utterances

We estimated the SLAM parameters for the subjects in the CSLU ERPA ADOS corpus using the regression model we created (see Section 6.2.2). We extracted the F0 curves from the speech recordings, calculated the features listed in Section 6.2.1 and estimated the one set of SLAM parameters for each child from all four diagnostic groups (TD, ALN, ALI, SLI). Finally, we examined whether there are significant differences between the diagnostic groups in terms of these prosodic features.

Data analysis

We performed four comparisons: ALN–TD and ALI–SLI, matching the groups on age, verbal IQ, and performance IQ; and ALI–ALN and SLI–TD, matching the groups on age (see details on how the matching was done in Section 4.2.4). We used a subset of the CSLU ERPA ADOS Corpus for this work (see characteristics in Table 6.3).

Table 6.3: The subset of the CSLU ERPA ADOS Corpus used for estimating the SLAM intonation model parameters

DX	n	age (mean, range)	amount of speech (mean, range; in seconds)
ALN	14	6.4 (4.7–8.2)	419 (164–980)
ALI	25	6.4 (4.0–8.8)	405 (45–1191)
SLI	14	6.4 (4.2–8.2)	451 (111–1167)
TD	28	6.0 (4.0–8.5)	524 (228–938)

6.2.3 Functional Data Analysis for prosody curves

We used FDA (see Section 3.3) for analyzing the intonation curves for 81 children from the CSLU ERPA Autism Corpus. FDA and specifically fPCA makes it possible to identify common components of a large set of prosody curves, and to represent those curves as a weighted sum of component curves. Inasmuch as the weights for the component curves show systematic differences between groups of speakers, it helps to differentiate between those groups automatically and to understand how the prosodic characteristics of those groups differ. In this section, we review our methodology for applying this technique to spontaneous speech samples to be able to examine the effect of the HFA and LI status on the intonation curves.

Converting F0 Curves to a Functional Basis

To make the F0 curves suitable for FDA, we first converted them to a sequence of B-splines, similarly to the procedure described by Gubian et al. (2010), making use of the R package called `fda` (Ramsay et al., 2014) (see also Section 3.3.2). First, we converted the F0 values to semitones, subtracted the mean F0 from each utterance, also dividing it by the standard deviation which may differ significantly between groups, to eliminate its effect. We made the curves continuous, which is a pre-requisite of the method, filling in the unvoiced parts by connecting their neighboring voiced samples with straight line segments. The method allows us to specify weights for the samples to regulate which ones are fitted most closely. We used the RMS value as the fitting weight for the voiced samples, since F0 values for louder voiced frames are generally considered more reliable, and a very small constant weight for unvoiced parts, so that the way we filled in the unvoiced parts should not matter much. We used a roughness penalty $\lambda = 10^{-2}$, which seemed to be suitable after visual inspection of the fit. (We could have used Generalized Cross-Validation to find the λ value for a good fit, but there is no theoretical guarantee that it would be the most suitable one, whereas visual inspection is important in every case.) We used B-Splines of order four, which results in curves with a smooth second derivative even at the knots between the spline segments (also called breaks), with one knot used for each F0 sample. Since the F0 curves need to be of the same length for fPCA, we fit each curve separately (i.e., using different number of B-splines); then resampled these continuous curves at 1000 points, which we fit again (with the same number of splines), this time without smoothing.

Functional PCA

We performed fPCA on the functional F0 curves for our diagnostic comparisons, calculating the first 10 eigenfunctions. We examined utterances with a length between 1 and 2 seconds, as utterances from a relatively short range are inherently similar and better aligned, and the bulk of the utterances was in this range. We oversampled the utterances such that the groups had equal numbers of utterances, so that both comparison groups contribute equally to the mean curve. Without this step, we might get higher eigenvalues for the group with fewer samples just because its mean is underrepresented in the common mean curve, which could skew the results. We excluded outlier utterances from our analysis, namely those with fPCA coefficients farther than 3 standard deviations from the mean, as such extreme values are probably the result of F0 tracking errors.

Data analysis

We performed t-tests with FDR correction on the per-subject means of functions of each fPCA coefficients, namely (following the notation of Section 3.3):

$$c_{i,m}, |c_{i,m}|, \frac{1}{M}\sum_m c_{i,m}, \frac{1}{M}\sum_m |c_{i,m}|$$

giving $2 \cdot M + 2 = 22$ features per subject. We included the mean of the coefficients to examine the possibility that the total amount of deviations from the mean or the total amount of accents may distinguish the groups from each other.

We compared the per-subject means of the individual fPCA coefficients and their absolute values, both for the unrotated and the rotated eigenfunctions, using t-tests with FDR to compare the per-group means.

6.3 Results of the Acoustic Analyes

6.3.1 Statistical features of prosody

Average F0 histograms

We created average histograms for each diagnostic group in the corpus, similarly to Bonneh et al. (2011). We did this by first calculating a histogram for each child from all per-frame F0 values shifted by the speaker's mean, then averaging these histograms and adding back the mean of the means subtracted earlier. The diagnostic groups were matched together on several cognitive measures, as described in Section 4.2.4. See Figure 6.7 for a pairwise comparison of these average

pitch histograms between the diagnostic groups. We can see that the average histograms look different for the two comparisons involving the TD group, ALN–TD and SLI–TD, whereas we cannot see a large difference when comparing the atypical groups. Apparently, the difference between the TD group and the other groups was much larger than between any pair of the latter. Our numeric features captured this difference, as we shall see below.

Group differences in F0 statistics

Below we describe the group differences between statistical features of F0 for groups matched two different ways, and for F0 curves extracted using two different tracking methods. We describe both analyses because we published the results of the first one (Kiss et al., 2012), whereas the second one dealt with the improved data used throughout this dissertation. We will see that the setup of the matched groups had a large effect on the final results.

We extracted F0 contours using the Snack toolkit (Sjölander, 2006) and the ESPS method. For the first analysis, we used a frame shift of 10 ms, a frame size of 7.5 ms, which results in 133 Hz lower bound, and we set the upper bound to 600 Hz. For the second analysis, we used the parameters described in Section 3.1, with the main difference being that we automatically adjusted the F0 range to that determined for the particular speaker. In both cases, we tried to correct the F0 jumps automatically. We calculated features related to statistical moments of F0 and intensity to characterize their distribution and checked for group differences. Below we review the findings for both.

100–600 Hz range for group-pairs We compared the per-subject statistical features of F0 in Hz between matched group-pairs using t-tests, 50–50 Manova (see Section 3.4.3), and Monte Carlo tests, and report the differences (Kiss et al., 2012) that were reliably present in each case. The group-pairs were matched independently of each other on appropriate measures (see also Section 4.2.4): TD and ALN on age, VIQ, and PIQ; TD and SLI on age; SLI and ALI on age, VIQ, and PIQ; and ALN and ALI on age. For the ALN–TD comparison, mean as well as median and MAD were significantly higher ($p < 0.05$), kurtosis and skewness were significantly lower ($p < 0.05$) for ALN. For the SLI–TD comparison, SD and MAD were significantly higher for SLI, as was median but to a much lesser degree. All four properties of the distribution, that is the location, spread, asymmetry, and peakedness of the per-subject F0 distribution differed significantly after FDR for the ALN–TD comparison, whereas spread and asymmetry were significantly different (but asymmetry much less) for the SLI–TD comparison.

We compared statistics of the per-utterance features as well, either the basic or the robust

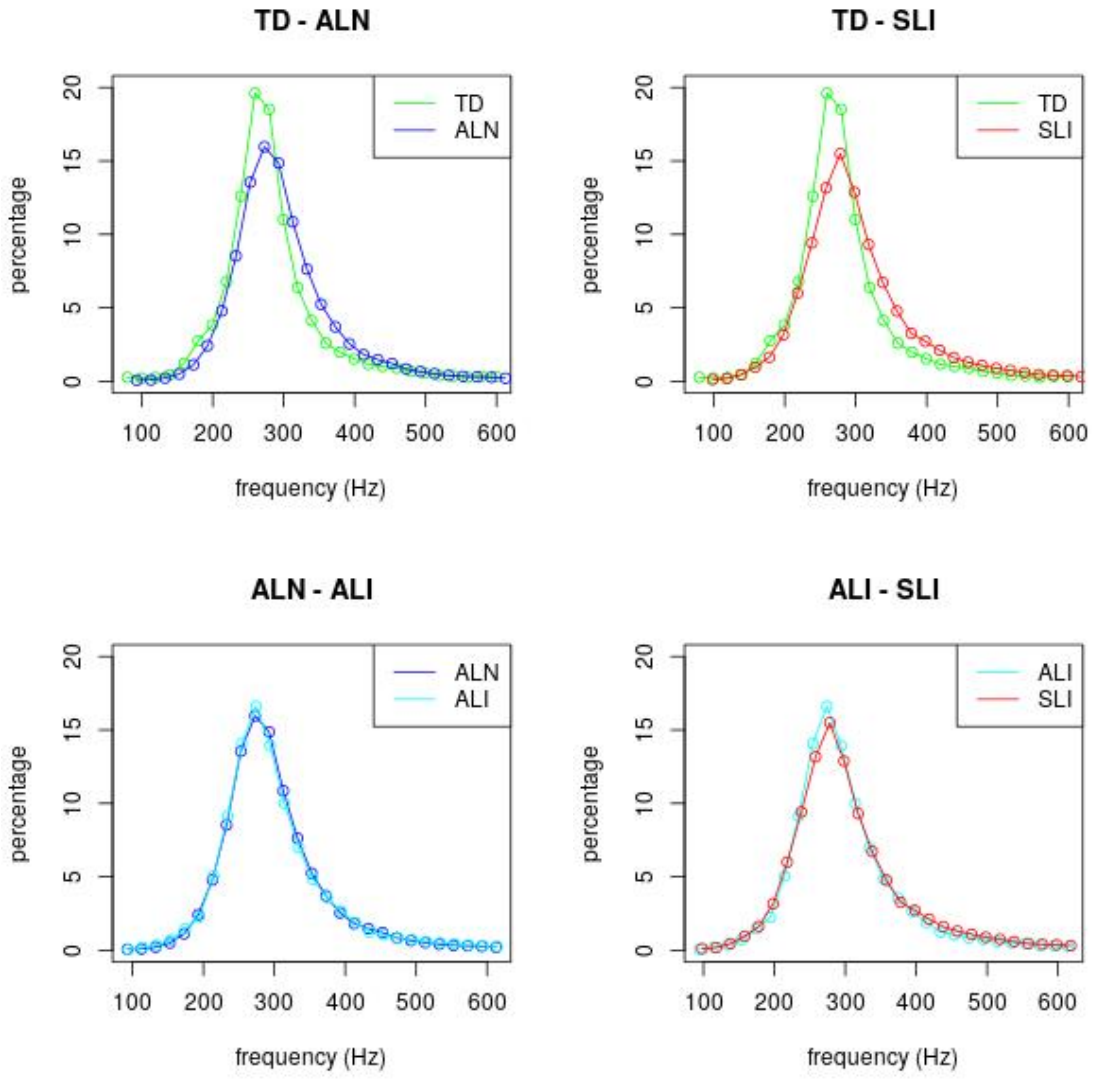


Figure 6.7: Average per-speaker F0 histograms for matched diagnostic group pairs (see Section 4.2.4). We calculated a histogram for each child from all per-frame F0 values shifted by the speaker's mean, averaged these histograms, and added back the mean of the means subtracted earlier.

versions. For ALN–TD, CoV of SD and of MAD, median of median and of MAD were significantly different ($p < 0.05$). For SLI–TD, MAD of mean, of median, and of MAD, as well as median of MAD were significantly different ($p < 0.05$).

Using this setup, we could replicate the findings reported by Sharda et al. (2010) for two of the three features they reported as significantly different for the ALN–TD comparison, namely mean of utterance means and ranges. For the SLI–TD comparison, all three features were significantly different, including the pitch excursion feature. We found the opposite for all but one measure reported by Bonneh et al. (2011): We also found the normalized F0 histogram peak height to be significantly different between TD and ALN as well as TD and SLI, but not F0 range or SD. However, we did find MAD, a robust measure of spread, to be significantly different.

Speaker-specific F0 range for groups matched together We compared the per-subject statistical features of F0 in both Hz and semitones between matched groups using Monte Carlo tests and FDR. The subjects for each group were the same in each group pair, as described in Section 4.2.4. We found much fewer significant differences than above, and only the robust statistical features seemed to give consistent results. For the ALN–TD comparison, we did not find significant differences in the per-speaker F0 statistics after FDR. For the SLI–TD comparison and F0 measured in Hz, the MAD–median ratio (a kind of robust coefficient of variation) was significantly different after FDR ($p < 0.05$).

We compared statistics of the per-utterance features as well, either the basic or the robust versions for F0 measured in Hz. For ALN–TD, median of MAD, of IQR, and of the MAD–median and IQR–median ratios were significantly different ($p < 0.01$) after FDR, as well as MAD of MAD and of the MAD–median ratio. For SLI–TD, median of MAD and of IQR were significantly different, as well as MAD of MAD ($p < 0.05$).

Regarding the features identified by Sharda et al. (2010) and Bonneh et al. (2011), for TD–ALN, we could replicate only the finding for the normalized F0 histogram peak height difference ($p < 0.01$). For TD–SLI, the same difference was present, as well as a difference in the pitch excursion feature ($p < 0.05$). There were no significant differences between the atypical groups after FDR.

Speaking rate

We compared the speaking rate between the diagnostic groups using all available utterances. A Monte Carlo test showed a significant difference between the four groups ($p < 0.02$). Further Monte Carlo tests for the matched group-pairs were all significant, expect for SLI–ALI ($p > 0.53$),

ALN–ALI ($p > 0.22$), and TD–ALN ($p > 0.13$). However, the merged HFA group (ALN and ALI together) spoke significantly slower than the TD group ($p < 0.011$), just as the merged language impaired groups (SLI and ALI together; $p < 0.002$).

We also analyzed the speaking rate using a mixed effect linear model with the subject id as the random effect and an intercept and a random slope for activity. The categorical fixed effects were the DX, sentence type, and activity, as well as whether the previous turn belonged to the examiner, and what its sentence type was. As numeric predictors, we used these sentence features: the duration (in seconds) as well as the number of pause fillers, mazes, and grammatical errors (as proxies for hesitation and thinking processes). Leaving the last three predictors out of the model did not change the results substantially. We measured the speaking rate in syllables per second (syls/sec), deriving it from the total utterance duration and the number of syllables, both determined from the transcripts. We excluded utterances with the most extreme values (about 0.2% with the highest, possibly erroneous values).

The marginal and conditional coefficient of variation for the model were 0.082 and 0.163, respectively. Since this model explains only a small proportion of the variance, finding a fixed effect that has a significant influence on the dependent variable does not mean that its influence is substantial, but it does mean that it has a measurable influence. Regarding the goodness of the fit, Pearson's r (product-moment correlation) between the observed and the fitted values is approximately 0.41, a moderate correlation. The estimates for the contribution of these fixed effects with their 95% confidence intervals are included in Figure 6.8; see the abbreviations of sentence types in Section 4.2.5. The intercept was calculated for TD statement in Conversation after the child's own statement (for DX, sentence type, activity, previous speaker, and previous sentence type); everything else is compared to the utterances with these properties.

The effect of all predictors was significant. The ALN, ALI, and SLI groups did not differ from each other significantly. The coefficients for ALI and SLI were significantly below zero ($p < 0.007$), whereas for ALN, the difference was not significant ($p < 0.10$; the 95% confidence interval is $[-0.320, 0.022]$).

The other predictors deserve to be discussed as well, if for no other reason but to show that the model results make sense. As we can see in Figure 6.8, the examiner generally interrupts utterances that are much slower than the average, and the child is also more likely to abandon such utterances. When the child interrupts the examiner, this interjection is relatively fast. Questions are generally faster than statements, whereas the utterance after the examiner's turn and especially those following yes–no questions are slower, probably because of the extra thinking necessary to answer. All activities have a slower average speaking rate than conversation. When children speak

faster, they tend to make more grammatical or word errors and produce more mazes (such as false starts and revisions). The presence of pause fillers is associated with significantly slower speaking rate as well. None of the above observations are really surprising, which shows that the model decomposes the phenomenon under discussion realistically.

Sentence-internal pauses

We analyzed the number of sentence-internal pauses using a mixed effects linear model. The explanatory variables were the number of words in the sentence, the DX, sentence type, activity, and their interactions. We included the subject id as the random effect. The number of words is naturally an important predictor, and their effect is significant even after FDR. There were significantly more internal pauses in abandoned sentences and during the play situation. There was also a significant interaction between activity and diagnosis for autism: the children in the autism groups had relatively fewer internal pauses during play. We did not find a difference in the correlation between F0 range and mean syllable durations, but the ratio of F0 standard deviation to the syllable duration standard deviation is significantly greater than zero for the language impaired groups, whereas it is not for the groups with normal language. We did not find a significant difference between the diagnostic groups.

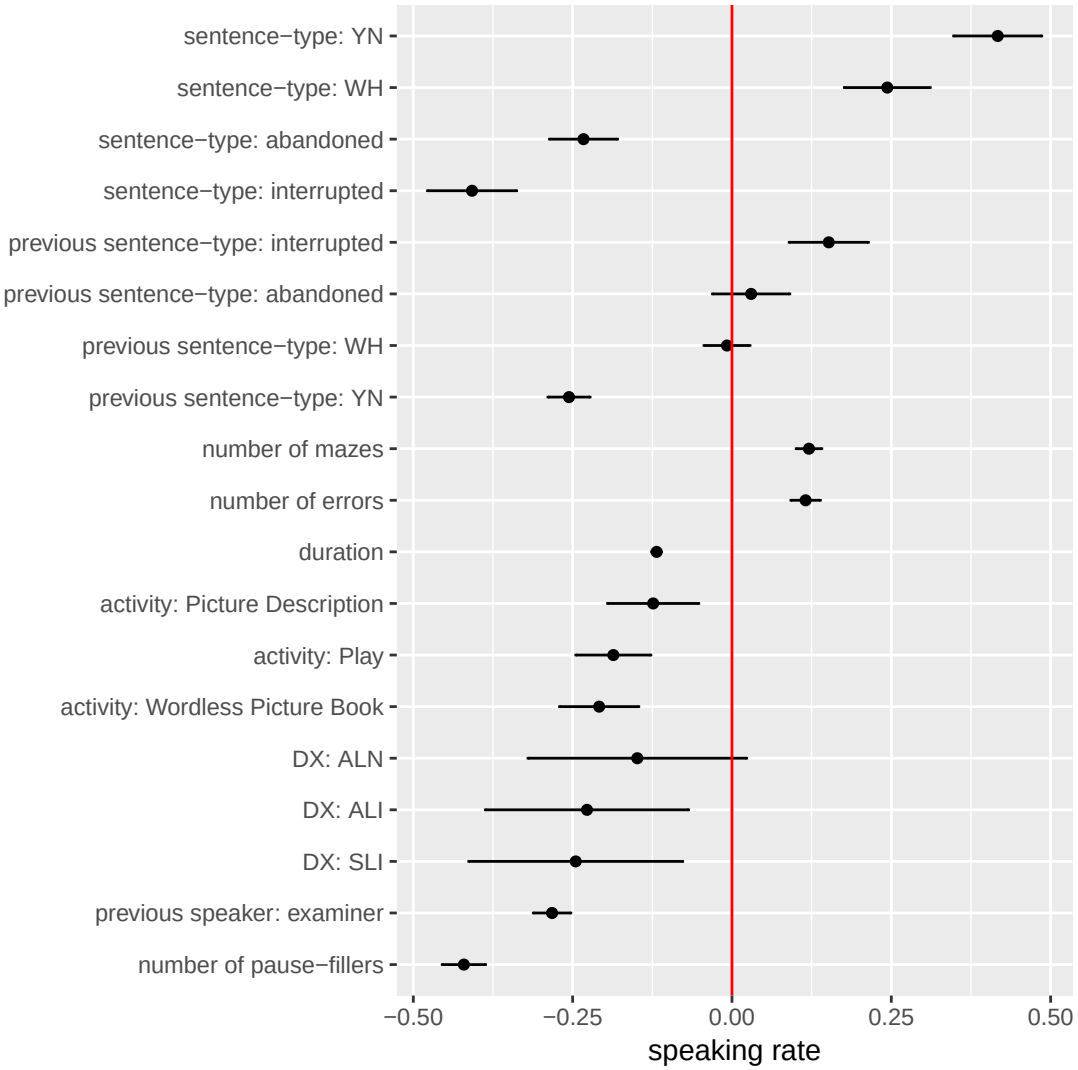


Figure 6.8: Speaking rate (in seconds per syllable): confidence intervals for the coefficients for the fixed effects of a MELM model. The intercept corresponds to TD statement in Conversation after the child's own statement (for DX, sentence type, activity, previous speaker, and previous sentence type). The sentence types YN and WH are yes–no questions and wh-questions, respectively.

6.3.2 Speaker-specific intonation model parameters

We estimated one set of SLAM intonation model parameters for each subject in the corpus and compared these across the groups. The phrase start and inflection point values differed significantly in comparison to the TD group (see Figure 6.9):

- ALN vs. TD: $p < 0.03$ for Manova on all six parameters; $p < 0.02$ for phrase start, and $p < 0.021$ for inflection point
- SLI vs. TD: $p < 0.04$ for Manova on all six parameters; $p < 0.006$ for phrase start, and $p < 0.01$ for inflection point
- The ALN–ALI and SLI–ALI group-pairs did not differ from each other significantly.

None of the other phrase and accent curve parameters differed significantly. In the case of the accent curves, the reason may be that our estimates scatter quite a lot around the actual values, as we saw in the evaluation of the regression model.

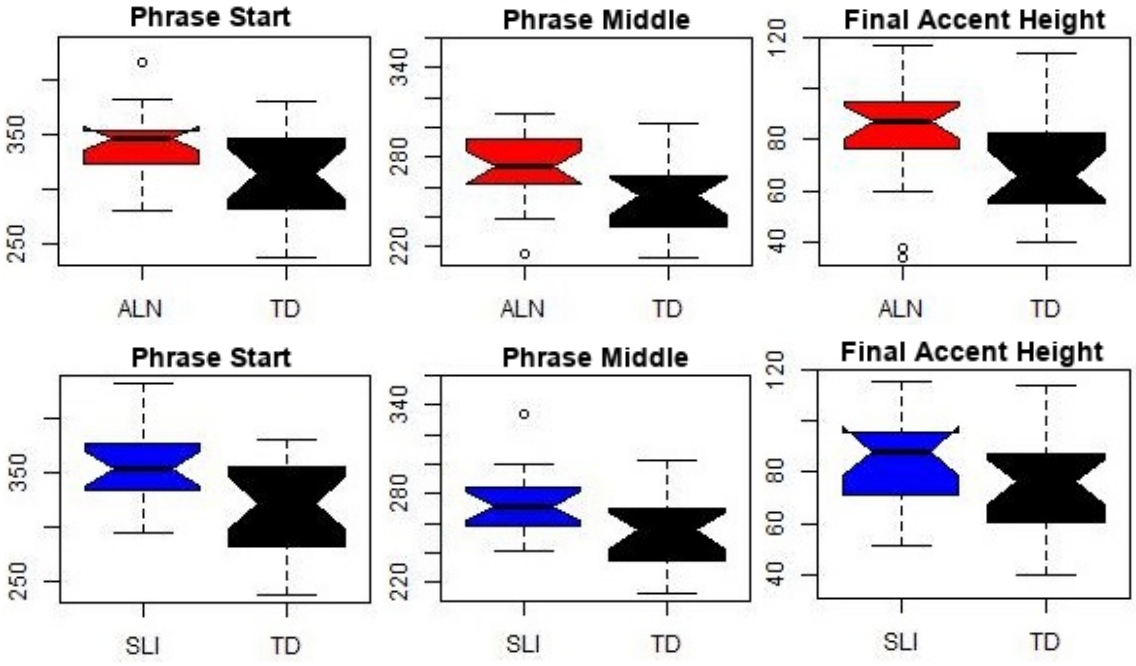


Figure 6.9: Estimated speaker-specific SLAM parameters (one value per subject) in Hz for matched diagnostic group pairs. We are not showing the ALI–SLI comparison because the difference was not significant.

6.3.3 Functional Data Analysis for prosody curves

We calculated the fPCA coefficients for the utterances of each subject in the corpus using both the unrotated eigenfunctions and the varimax rotated ones and compared them across the groups. For an illustration of the eigenfunctions (i.e., curve components) corresponding to each fPCA coefficient, see Figures 6.10 and 6.11; the curve in the middle is the mean curve, the ones formed with plus and minus signs are the mean plus or minus the eigenfunction multiplied with one standard deviation of the corresponding coefficient. The two sets of eigenfunctions represent complementary aspects of the curves: Based on the visualization of the eigenfunctions, the unrotated ones correspond to a frequency analysis of the peaks, while the varimax rotated ones show differences in peak heights at various positions of the utterances.

The fPCA coefficients for the unrotated eigenfunctions differed significantly between TD and the other diagnostic groups, but not among the atypical groups. Namely, for utterances between 1 and 2 secs, the mean of coefficients 4 and 6 were significantly different after FDR for the ALN–TD ($p < 0.036$) and SLI–TD comparisons ($p < 0.006$), with smaller (negative) values for the TD group (see Figures 6.12 and 6.13). The difference was even greater when we compared TD to the merged HFA (i.e., ALN + ALI) and LI (i.e., SLI + ALI) groups ($p < 0.003$ and $p < 0.001$, respectively).

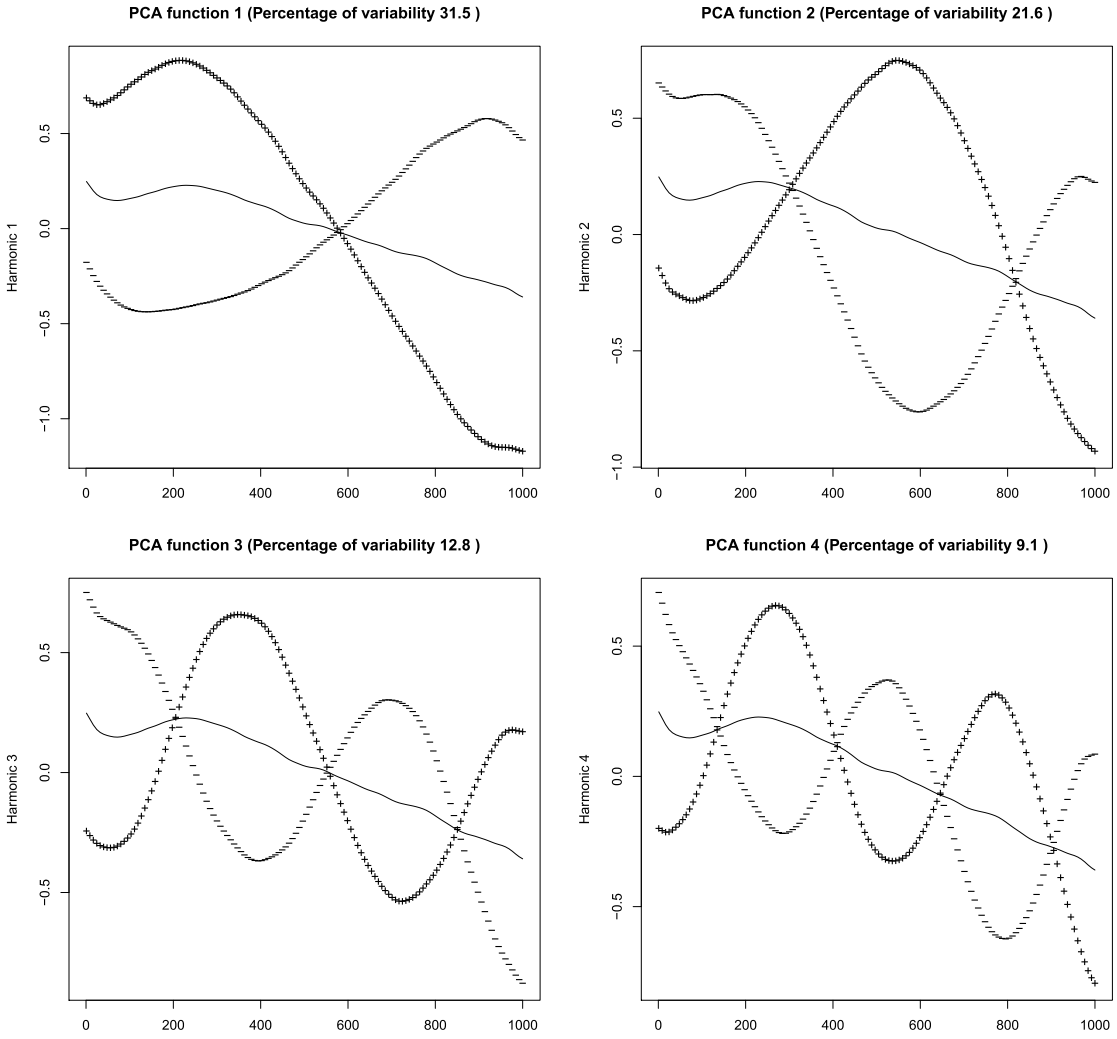


Figure 6.10: The first four of 10 unrotated eigenfunctions with 1000 B-Splines for utterances between 1 and 2 seconds for the ALN–TD comparison. The curve in the middle is the mean curve, the ones formed with plus and minus signs are the mean plus or minus the eigenfunction multiplied with one standard deviation of the corresponding coefficient.

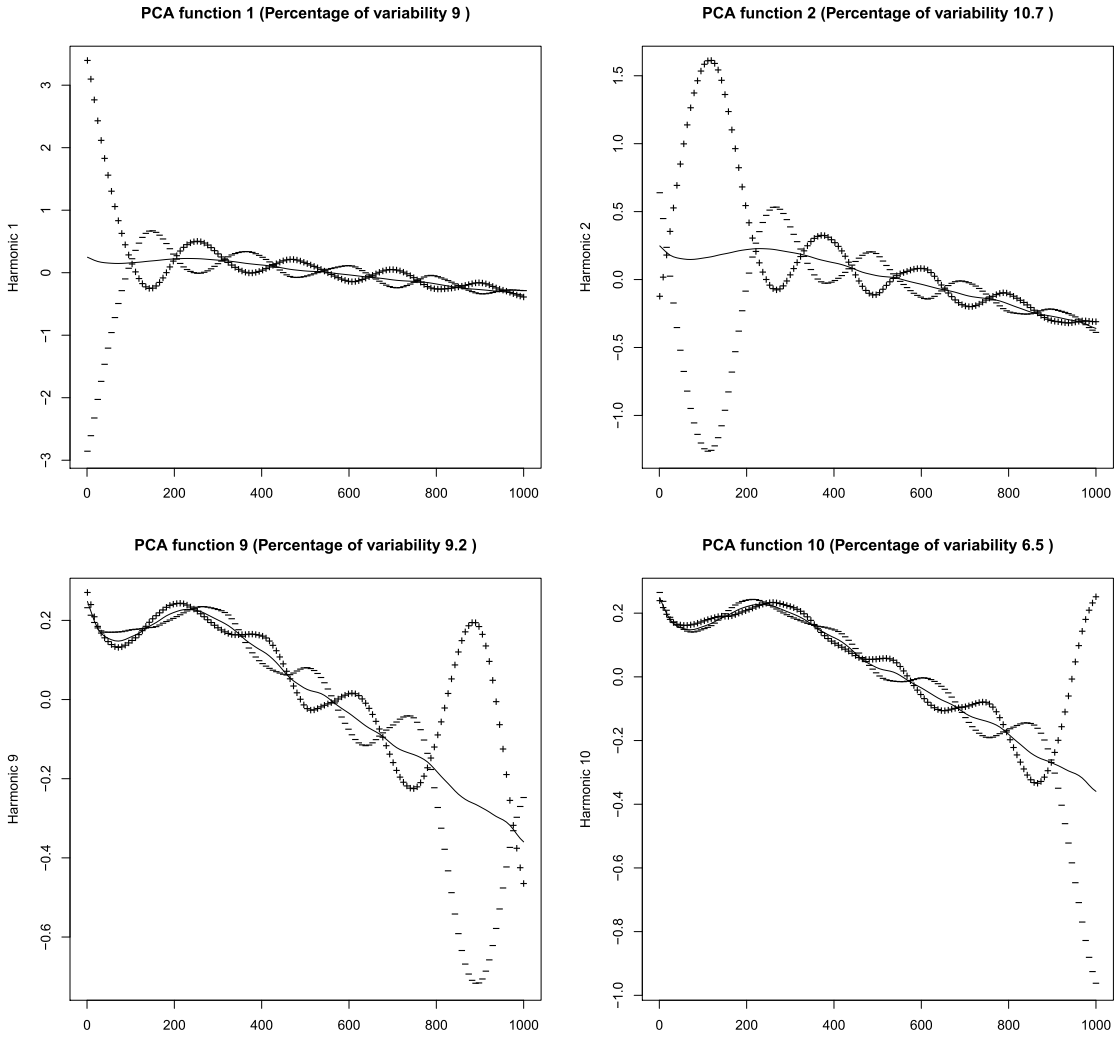
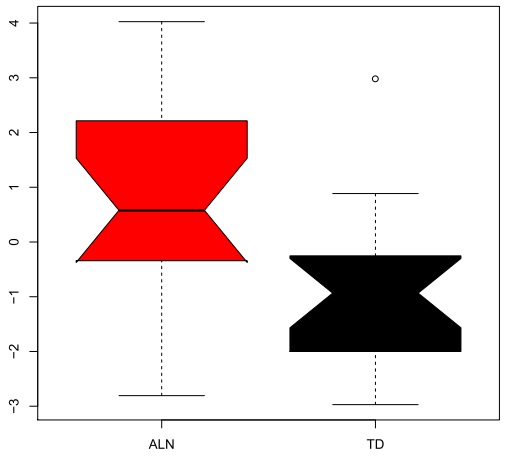
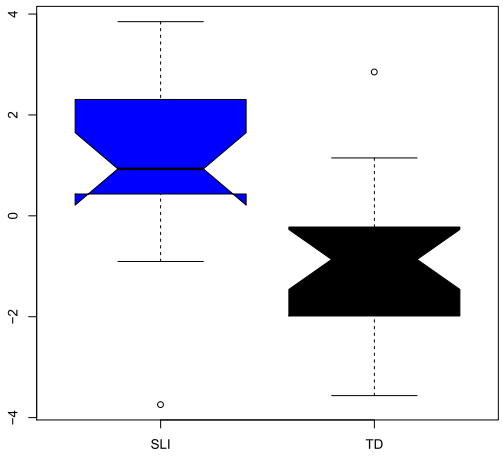


Figure 6.11: The first, second, ninth, and tenth rotated eigenfunctions with 1000 B-Splines for utterances between 1 and 2 seconds for the ALN–TD comparison. The curve in the middle is the mean curve, the ones formed with plus and minus signs are the mean plus or minus the eigenfunction multiplied with one standard deviation of the corresponding coefficient.

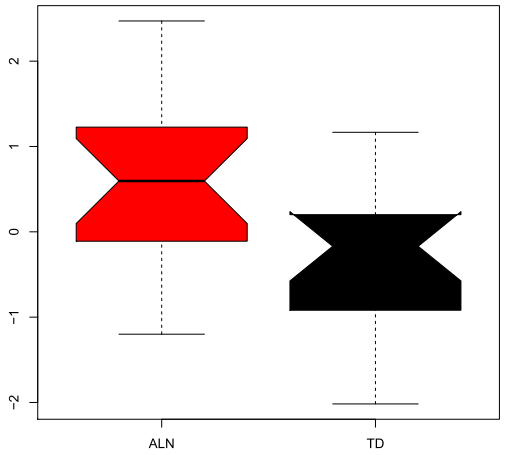
One coefficient for the rotated eigenfunctions differed significantly after FDR: Coefficient 10 is smaller for SLI than for TD ($p < 0.022$, uncorrected $p < 0.001$); see Figure 6.14. For ALN–TD, the difference is not significant after FDR, but it is for the merged groups (HFA–TD: $p < 0.0011$; LI–TD: $p < 0.0001$).



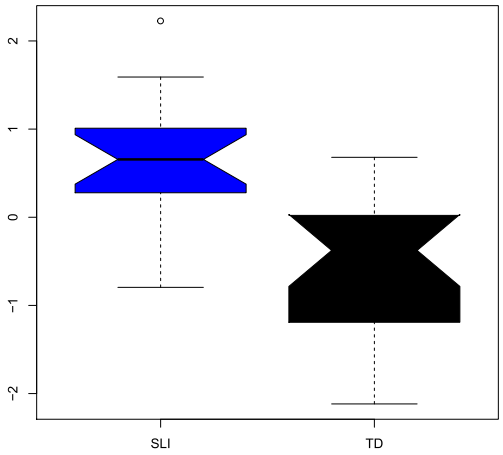
(a) Unrotated fPCA coef. 4, ALN vs. TD



(b) Unrotated fPCA coef. 4, SLI vs. TD

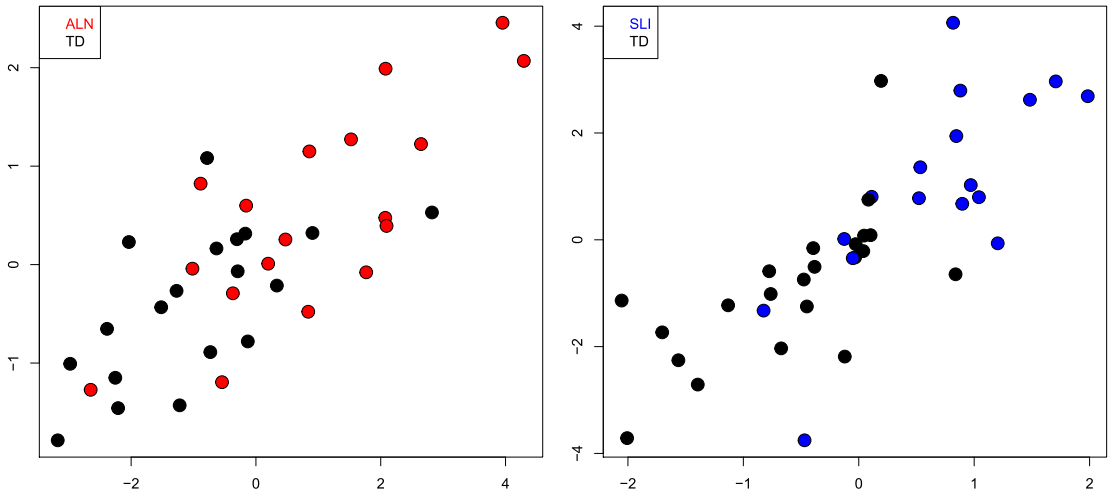


(c) Unrotated fPCA coef. 6, ALN vs. TD



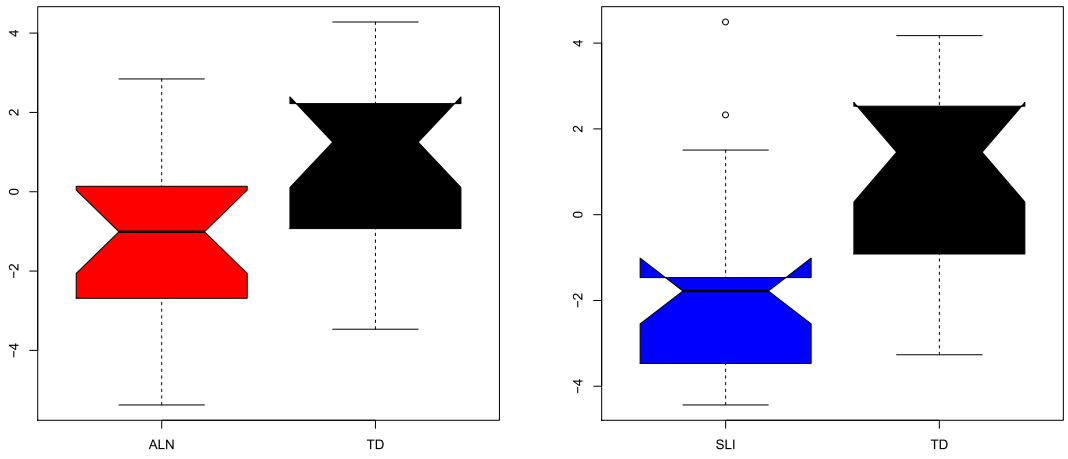
(d) Unrotated fPCA coef. 6, SLI vs. TD

Figure 6.12: Boxplots of the per-subject means of unrotated fPCA coefficients 4 and 6 for utterances between 1 and 2 seconds. The difference was significant after FDR for the ALN–TD ($p < 0.036$) and SLI–TD comparisons ($p < 0.006$).



(a) Unrotated fPCA coef. 4 vs. 6, ALN vs. TD (b) Unrotated fPCA coef. 6 vs. 4, SLI vs. TD

Figure 6.13: Scatterplots of the per-subject means of unrotated fPCA coefficients 4 and 6 for utterances between 1 and 2 seconds. The difference was significant after FDR for the ALN–TD ($p < 0.036$) and SLI–TD comparisons ($p < 0.006$).



(a) Rotated fPCA coef. 10, ALN vs. TD (b) Rotated fPCA coef. 10, SLI vs. TD

Figure 6.14: Boxplots of the per-subject means of rotated fPCA coefficient 10 for utterances between 1 and 2 seconds. It is smaller for SLI than for TD ($p < 0.022$, uncorrected $p < 0.001$). For ALN–TD, the difference is not significant after FDR.

6.4 Discussion of the Acoustic Differences

We acquired results using a diverse set of methodologies. These are on the one hand complementary to each other, on the other hand, the results seem to converge into the same direction. First, we discuss the results separately for each approach, then we summarize and interpret our findings.

6.4.1 Statistical features of prosody

The apparent differences in the average F0 distributions between the diagnostic groups and the corresponding differences in the summary statistics are relevant from at least two viewpoints. First, it indicates that such statistical properties of F0 are suitable for being used as features for machine learning. Second, these differences shed some light on qualitative differences in the intonation curves of children with HFA and SLI. What these values definitely indicate is that the F0 of children with HFA is on average more variable, both within the utterances and across utterances for the same speaker. The differences in the medians of per-utterance statistics indicate more within-utterance spread in HFA and SLI, that is, the F0 contours are less flat than in TD. The spread of per-utterance spread also shows more variability across utterances in HFA and SLI: F0 is less repetitive than in TD. However, based on these findings only several explanations are possible regarding how the intonation curve shapes differ, as the location information of F0 peaks and valleys is not utilized in any way in the calculations of these statistics.

Robust statistics have proved to be more useful features than their non-robust version. For example, the robust measures of location and spread (median, MAD) of the per-subject F0 values in Hz distinguished the groups even though the non-robust measures (mean, SD) did not. Outliers occurring due to tracker errors obviously cause the latter to digress in unpredictable ways.

Several other studies have identified some of the same statistical F0 features as beeing different between ASD and TD, as well as different ones. Sharda et al. (2010) reported significant F0 differences based on their conversational speech data in the mean, range, and per-utterance pitch excursion. On our data, depending on the matching approach, either we could replicate all of them or none of them. So the results were not reliable: They depended on the particular subjects being included even when the high-level matching approach was identical. Bonneh et al. (2011) reported significant pitch differences in SD, range, and the logarithm of the highest peak of the F0 histogram, but no significant difference in mean, using the speech collected during a picture-naming task. We could replicate consistently only the finding regarding the F0 histogram peak, but that feature indeed seems to be the most robust one among all statistical features examined.

Note that since we used conservative significance tests, when a feature is significantly different

for only one of the ALN–TD and SLI–TD comparisons, it does not mean that ALN and SLI are significantly different on that measure. In fact, the atypical groups were very similar regarding all our measures. As we can see, the distribution of F0 values is atypical not just in HFA, but also in SLI. Moreover, we did not find any significant differences after FDR for any of the pitch feature sets for the SLI–ALI and ALN–ALI group comparisons.

The language impaired groups had a significantly lower speaking rate than the TD group. This difference persisted even after controlling for content features: It is not due to differences in the utterance lengths, the number of pause fillers, mazes, errors, the type of examiner utterances, or the ratio of utterances from various activities and with various sentence types. The average speaking rate for the ALN group was mathematically lower compared to TD, but this difference did not reach significance ($p < 0.10$). It may, however, become significant when more subjects become available, although likely with a smaller effect size than for the TD–LI comparison. There is a considerable variation in the average speaking between the subjects in each group. In fact, the fastest speakers have approximately the same average speaking rate in each group (around 3.4 syls/sec), and the slowest speakers are not much slower in the groups with disorders (around 2.0 syls/sec) than those in the TD group (around 2.4 syls/sec).

Bonneh et al. (2011) reported significant differences between their ASD and TD groups on a picture-naming task. They worked with 41 Jewish children with ASD and 42 TD controls from a similar age range (4 to 6.5 years). All children spoke out the Hebrew names of the same set of daily life pictures. (The authors regularly use the word "reading" for this, but both the age range of the children and the task description indicates that the pictures did not have labels; for example, they say that the examiner triggered the responses by asking the question "What's that?".) They found a slower speaking rate (measured in number of words per minutes) in their ASD group compared to their TD group, but they did not present information on the language competence of their ASD subjects. We found a similar difference in our data between the children with ASD (44 subjects) and those with TD (31 subjects), but not between ALN (19 subjects) and TD. In other words, once we removed the children with language impairment from the ASD group, the difference was no longer significant. This may mean that it is having language impairment that contributes to the lower speaking rate they observed, not autism per se, or at least the effect of language impairment is larger. But we cannot state that decisively. The main reason is that the two stimulus sets were very different: we analyzed a larger amount of spontaneous speech, whereas Bonneh et al. used one minute of read words. Secondly, excluding the language impaired children from our ASD group reduced the number of subjects substantially thus reducing the power of our statistical test.

6.4.2 Speaker-specific intonation model parameters

The SLAM intonation model phrase start and phrase middle (or inflection point) parameter estimates for the ALN and SLI groups differed significantly from that of the TD group, but not in the phrase end parameter. This indicates a qualitative difference in the average shape of the intonation curves of the children in the groups with disorders. Namely, their utterances generally start at a higher pitch, and arrive at around the same final pitch value. We did not find significant differences in the accent curve height parameters, probably due to the fact that the accuracy of the regression model was relatively low for these parameters even on the artificial data. Nevertheless, the final accent height was on average higher in both the ALN and the SLI groups than in the TD group.

The success of this method depends on a number of assumptions: First, we assume that the speaker's prosody can be accurately characterized by the SLAM model. This is a reasonable assumption, as the model in its current form is able to cover a wide variety of intonation curves with the exception of parenthetic remarks, and we do not expect children of these ages to use parentheticals. Second, regarding the per-speaker estimate, we posit that the speaker uses the same kind of intonation (characterized by the same SLAM parameters) for all of his utterances. This may not be true. Moreover, in natural speech, the intonation can be affected by emotion, mood, and the particular lexical content. Nevertheless, it is reasonable to assume that the determined parameters are characteristic of the speaker, with the instances for the individual utterances varying around it. Third, the speaker's parameters must be in the range used for training the model. We made efforts to ensure this. Fourth, the F0 curves are correctly detected from the speech signal. An obvious solution would be to use hand-corrected data, but it is not currently available to us. We strived to ensure that the tracker output would not be wrought with much error. Moreover, the fact that we only used robust statistics of F0 is aimed at alleviating the issue of tracking errors. Fifth, we reckon that there is a unique mapping between a sentence curve and the corresponding SLAM parameters. This is generally true, except for utterances that start with a non-sonorant part. In the latter case, the balance between the height of the phrase curve and the accent curve is not obvious. Otherwise, the specific choice of interpolation functions in SLAM (namely, linear segments and cosine interpolation) makes the relationship uniquely determined. Finally, the estimate can only be as good as the SLAM implementation we use. We made efforts to validate the correctness of the particular implementation we worked with.

Based on the above considerations and the extensive validation on the artificial data, it is reasonable to believe that this model yields reasonable estimates of the average speaker-specific

intonation model parameters for spontaneous speech corpora. We saw that the model yielded reasonably good estimates for the phrase curve parameters during validation despite the difficulty of the task. Note that we generated the synthetic training data in a way that resembles the real-life situation in that the utterance contents differed for each artificial speaker. We deduce from this that the variation of the parameter estimates around the actual value for real speakers is probably not substantially larger than what we saw for the artificial data.

6.4.3 Functional Data Analysis for prosody curves

The differences in fPCA coefficients between the diagnostic groups can be useful for automatic classification or regression, but it is not immediately obvious how they correspond to overall differences in the intonation curve shapes. The result from the unrotated coefficients, namely the fPCA coefficients 4 and 6 being lower for the TD group, seems harder to interpret. It may indicate a higher variability in the utterance curve, and also agrees with the other finding of a higher final F0 value in the TD group.

The result of fPCA for the rotated coefficients shows that the average final F0 value is significantly higher for the TD group for the utterances analyzed, namely those between lengths 1 and 2 seconds, which is the bulk of the utterances. This difference is not attributable to differences in the mean or spread of F0 since we performed this analysis on the standardized values. A possible interpretation is that this is the consequence of a difference in the speaker-specific intonation that the estimated SLAM intonation model parameters have shown: We found higher average utterance start and middle values in statements for HFA, but no significant difference in the minimum F0. This results in a higher utterance mean F0 value. If we subtract this mean F0 value from each utterance, as we did for the FDA analysis, the utterance-final values will be shifted lower when the utterance start is higher, which is exactly what we found.

There is room to improve our FDA analysis, for example by using Landmark Registration (Ramsay et al., 2009): By aligning important points within the utterances, we add knowledge to the task at hand, and can thus make sure that important differences in the curves are not smeared out. The landmarks can be, for example, F0 peak locations, phoneme or syllable boundaries. This approach requires analyzing together only a subset of the curves, with an identical number of these landmarks. This reduction in the number of utterances may not hurt our ability to reach significant results, as the differences themselves can be more pronounced. For example, we got more significant differences when working with only the short utterances than for all the utterances, probably because they are inherently better aligned and more similar to each other.

6.4.4 Summary

Each approach yielded significant differences between the TD group and the atypical groups. Just as consistently, we were not able to identify significant differences between the groups with HFA and SLI. This means that having either HFA or SLI affects the prosody, but a combined diagnosis of the two does not make it more atypical, at least to the extent that our acoustic features can capture this difference. Nonetheless, some statistical features of F0 were significantly different for the HFA–TD, but not for the SLI–TD comparison, so some differences may become observable between the HFA and SLI groups for an even larger data set.

These results shows that we must be careful when interpreting findings where only a TD control group is used. Some of the atypical features may not be specific to autism but instead may be part of the phenotype of several disorders, may indicate comorbidity, or may be a characteristic of a subtype of ASD. Whether or not listeners perceive the utterance-level prosody of children with SLI as having atypical prosody, similarly to ASD (Nadig and Shaw, 2012), is an important question, with which we deal in Chapter 7.

6.5 Summary of the Acoustic Analyses of Prosody

We presented a study on the atypicality of pitch in neurodevelopmental disorders, namely HFA and SLI, using the CSLU ERPA ADOS Corpus, a relatively large and very well characterized database. We compared statistical F0 features, per-speaker estimates of the parameters of the SLAM intonation model, and fPCA coefficients. We found significant differences compared to the TD control group in the per-subject pitch distribution, as well as statistics of sentence-level features, but did not find significant differences between the HFA and SLI subgroups.

We proposed a method to estimate speaker-specific intonation parameters in the SLAM (van Santen et al., 2004) intonation model. The regression model can be trained using synthetic data only, and can be used for spontaneous speech and requires the utterance boundary markup only. Extensive evaluation of the model indicated that the estimates for the intonation curve parameters are reasonably acurate; the current feature set however did not provide satisfactory estimates for the accent parameters. It is also proof that overall statistics can be used to draw inferences about individual pitch contours, through estimating the parameters of the SLAM intonation model. The approach yielded useful estimates for spontaneous speech that differentiated typical speech from the speech of children with neurodevelopmental disorders. Specifically, children with HFA and SLI have significantly higher phrase start and middle F0 values on average.

We quantified atypical aspects of the intonation curves using Functional Principal Component

Analysis. We found significant differences compared to the typically developing group, but did not find significant differences between the diagnostic groups.

One of the consistent findings using any of the approaches was that the prosody of children with HFA differed from those with TD, but not from those with SLI. This confirms that atypical prosody is not specific to ASD.

Chapter 7

Perceptual Ratings of Prosody: Collection and Analysis

In the previous chapter, we delved into prosody in autism through the exclusive use of objective features and computational techniques. Now it is time to put the human observer into the loop. In this chapter, we examine prosody through the perceptual judgments of naive listeners.

7.1 Motivation for Collecting Perceptual Atypicality Ratings

As we have seen earlier in Section 2.3, prosody in autism is known to sound atypical, but it is not known in detail how it differs from typical speech, and how that prosodic difference relates to the contents of the speech. Based on our review of perceptual studies in Section 2.3.3, children with ASD often perform significantly worse than their TD peers on several expressive and receptive prosodic tasks, for example in PEPS-C studies. Moreover, even when their prosody is functional, it still often sounds atypical. PVSP studies, which ask about atypical prosodic aspects of one utterance at a time, found inadequate stress patterns, and occasionally some other atypical aspects. Other perceptual studies generally asked about atypicality ratings for relatively long, multi-utterance speech segments, unmatched for content, and found that speech in ASD sounds significantly more atypical. To our knowledge, no previous study examined the relationship of prosody in autism to the emotional charge and to content.

In this chapter, we describe our procedure for collecting atypicality ratings for various aspects of speech (textual content, emotional charge, and prosody), which complements earlier procedures in several ways. Unlike PEPS-C, it deals with spontaneous speech and concentrates on the perceptual aspects of speech. Similarly to the PVSP, we worked with individual utterances, but the questions use everyday language and are simple enough to be answered by naive listeners with little training, although it requires pooling ratings from several judges for each utterance for attaining reliable gold-standard ratings. It is a multi-step methodology for examining aspects of not only prosody but also the textual and emotional content. We used this procedure to collect ratings for utterances of the CSLU ERPA Corpus (see Section 4.2) and analyzed these to identify group differences.

Prosody can sound atypical because it does not sound appropriate in the context, or the prosody of a particular utterance is atypical even in isolation. We posit that a single utterance can sound atypical in five different ways. First, the prosody can be atypical independent of content: No matter what the content is, the prosody would not suit it (for example, due to highly varying rhythm or intensity, atypically high pitch, too loud, or strangled voice). Second, the speaker may say an affectively and pragmatically typical content in a manner that is atypical for that content (for example, using improper stress placement). Third, it can sound atypical when the

speaker uses typical prosody for affectively or pragmatically atypical content, saying something surprising as if it was ordinary (for example, speaking about a dreadful topic in a cheerful voice). Fourth, the utterance can be atypical both from a prosodic viewpoint as well as pragmatically or having atypical affective content. Fifth, the prosody can be inappropriate for the context, that is pragmatically incorrect (e.g. speaking to an adult using motherese). It is all the more important to distinguish between prosody and content, as both aspects have been implicated in autism and their relationship is not always clear. For example, Diehl et al. (2009) talk about "pedantic" speech, apparently referring to prosody, whereas Ghaziuddin and Gerstein (1996) characterized the use of pedantic speech, seemingly interpreting it primarily referring to content.

We hypothesize that the prosody of children with autism or language impairment sounds significantly more atypical to naive listeners than those of typically developing controls, at the utterance level independent of the context, even after controlling for the effect of content and emotional charge. To the best of our knowledge, no previous study examined exactly this question.

We acquired ratings about various aspects of speech utterances from naive raters who were agnostic to the diagnostic status of the children. We asked about the perceived emotional charge, whether the words or the meaning was unusual, and whether the prosody of their utterances sound unusual and in what way. We received feedback and advice from colleagues at CSLU in the process, among others from Alison Presmanes Hill, Éric Fombonne, Peter Heeman, as well as some JavaScript code examples for the rating interfaces from Rebecca Lunsford, for which we are grateful. Any faults remaining in this work are our own.

7.2 Methodology for the Perceptual Rating Collection

7.2.1 Study Design

We carefully designed the data collection to ensure content validity. Below we describe in detail the perceptual rating tasks, the stimuli that we used in those, the task interfaces, the raters and the process of the rating collection, and finally the data analysis and the planned comparisons. We conserved the source code for these steps, mostly in the internal R packages `GProsodyRatings` and `GAMTRatings`, and the collected ratings as R data in `GProsodyRatings`; one can find a short description for these in Appendix B.1. First, we summarize the rationale; then we relate the details in the following sections.

We created four tasks for obtaining information on various aspects of the children's speech, sometimes separating different modalities: the (textual) content only, the prosody only (rendering the speech unintelligible), and the original speech (content and prosody), either asking high-level

questions or requiring detailed ratings about atypical aspects. This design also allowed us to sequence the tasks such that we could use the results from earlier tasks to pre-filter the stimuli for later ones; for example, we asked detailed ratings for different prosodic components (such as pitch or rhythm) only for the utterances that the majority of raters perceived as sounding "atypical".

The task interfaces evolved during the process, and we used two main versions. While this can be viewed as a drawback, it also gave us the opportunity to estimate the effect of this change on the results. We describe those in detail below in Section 7.2.3.

We created two sets of stimuli (see more in Section 7.2.4): one containing utterance pairs matched on expected prosody from comparison group pairs (ALN–TD and ALI–SLI) based on the textual and emotional content (Matched on Expected Prosody, `MEP`) and another containing maximally diverse utterances for each individual, covering the prosodic repertoire of each subject to the extent possible using a limited number of utterances (Maximally Individually Diverse, `MID`). The `MEP` set helps us to identify prosodic differences between the groups that are not due to the differences in content through the use of utterance pairs that are as similar as possible in their content features. Note that this concept is similar to minimal pairs in phonology. The `MID` set on the other hand aids in showing overall group characteristics as well as in portraying the individuals making up the groups better. Naturally, one can combine all data for training regression models to estimate perceptual ratings.

We conducted the rating collection through a crowdsourcing website (Amazon's Mechanical Turk). The Institutional Review Board of the Oregon Health and Science University (OHSU) permitted the use of crowdsourcing websites for collecting perceptual ratings for the speech samples, with certain precautions. The children's conversations may contain Protected Identifiable Information, such as proper names, residential addresses, and ZIP Codes. Therefore we filtered out from consideration any utterances that seem to contain such information: We identified words that are potentially proper names based on surface features of the tokens (e.g. capitalization), eliminated false positives from this list manually, then we excluded the utterances that contain any of these words. We only allowed raters to do the tasks who are from the US and whose prior work has been accepted by the recruiters most of the time. We also monitored their work and evaluated their competence.

We aggregated the ratings to get per-utterance scores, taking into account our rater competence estimates as well. We analyzed the data using the methods outline at the end of this chapter, either making planned comparisons or applying appropriate FDR correction. Now let us see the details for each step of this procedure.

7.2.2 Tasks

As mentioned earlier, we created tasks for different modalities, asking about diverse aspects of the speech recordings. Below we review the four tasks we created, denoting them by their mnemonics in the subsection titles. We present the task interfaces in Section 7.2.3.

The `TEXT` task

This task is for ratings based on the textual content of the utterances only: We show the rater the transcript of the stimulus utterance, together with the approximate age and gender of the child. We ask questions about the emotional state of the speaker using the dimensional model of emotions (specifically, the arousal and valence scales), whether there are unusual words, and if the utterance overall has an unusual meaning.

The `SPEECH` task

We play the speech recording for the utterance, and additionally display to the rater the same information as during the `TEXT` task (the transcript and the approximate age and gender of the child). We ask for ratings of arousal and valence, whether the prosody sounds atypical, and whether the prosody and the text are congruent with each other.

The `DELEX` task

This task is very similar to `SPEECH`, only the text is not shown, and a delexicalized version of the speech is played to the rater, which is rendered unintelligible while preserving prosody to the extent possible. We used the delexicalization method developed by Alexander Kain and colleagues (Kain and van Santen, 2010), which they have shown to preserve prosody faithfully, along with naturalness, speaker identity, and emotions. We ask for ratings of arousal and valence, and whether the prosody sounds atypical. Since the rater does not know what the content is, we do not ask about agreement between prosody and text.

Note that this task is more ambiguous than the `SPEECH` task, because during delexicalization, some information is inevitably lost, for example the vowel durations. It is hard to tell for example if a short vowel is protracted, or a long vowel is shortened.

The `SPEECH ASPECTS` task

This task is similar in its purpose to the PVSP, as it is for identifying the aspects of prosody that are unusual. The difference in its use is that we use it only for samples that some raters have

already identified as having atypical prosody, or prosody that is incongruent with the contents. In theory, this approach reduces the amount of work to be done without losing any information. The raters are free to mark any number of the speech aspects (including none) as atypical. We ask them about seven aspects of prosody:

1) stress

2) intonation

3) pausing

4) speed

5) loudness; and

6) voice quality.

7.2.3 Task interfaces

We first created a set of simple task interfaces (denoted `S` in tables and figures), one for each task described earlier. These employed yes / no questions, occasionally 3-point scales, and 5-point scales for the emotional arousal and valence ratings. For all four tasks, we used it to collect information for the `MEP` stimulus set. Later colleagues who helped review the tasks raised concerns about the validity of instructions and the sufficiency of the rating scales. Their feedback prompted us to change the instructions and interface, taking into account feedback obtained through Amazon Mechanical Turk as well. The result is called the detailed task interfaces (denoted `D`). Below we summarize the task interfaces and the rationale for each one.

Simple task interfaces (`S`)

We had to decide how to ask about emotions in a way that is both meaningful to us and relatively easy to explain to naive raters. One can either use emotion labels (such as happy, fearful, angry, sad) or the dimensional model that represents emotions as points in a multidimensional space. Psychological studies have identified multiple dimensions that can adequately represent emotional states (see e.g. Fontaine et al., 2007), with two of the most important dimensions being *valence* (how negative or positive the feeling is) and *arousal* (the amount of physical response; only third in importance after *control*, yet more often used in similar studies). We decided to use the dimensional approach because it allows us to measure the intensity of the emotions directly while also enabling us to project these ratings to emotion categories if necessary (see e.g. Figure A.1). We tried to

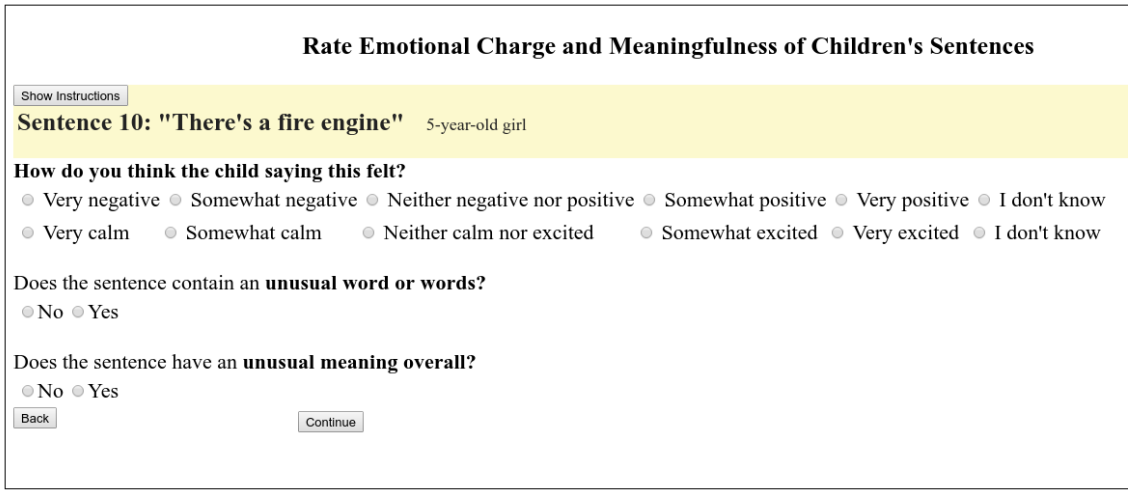


Figure 7.1: Simple (`S`) task interface for the `TEXT` rating task

explain the concepts of valence and arousal to raters using everyday words, also clarifying that arousal does not refer to the intensity of the emotion. The words "negative vs. positive" for valence and "calm" and "excited" for arousal seemed suitable. We used a five-point scale and allowed the raters to choose "I don't know." They may not be able to identify for example arousal based on the textual content. The complete instructions and questions for the simple task interfaces are available in Appendix A.1.

We asked about the atypicality of the content and the prosody using a discrete scale. These can be considered to lie on a continuum, but using a two or three-point scale seemed enough. Direct Magnitude Estimation (DME) would be more appropriate for rating atypicality if we wanted to get ratings on a continuum, or a fine scale (Campbell and Dollaghan, 1992), but we do not believe that using DME would make a real difference for the discrete scale we used. We allowed the raters to listen to the audio as many times as they wanted to, and they were also shown the transcription of the sentences, giving them an opportunity to give feedback if the two did not agree (this did happen in a few cases). Figures 7.1 through 7.5 show some screenshots of the rating interfaces.

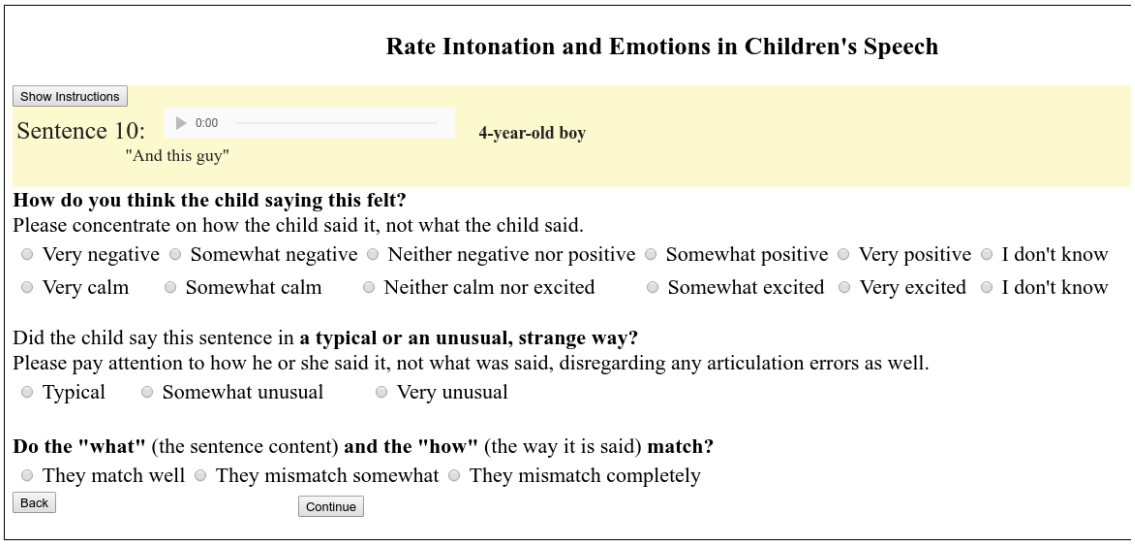


Figure 7.2: Simple (`S`) task interface for the `SPEECH` rating task

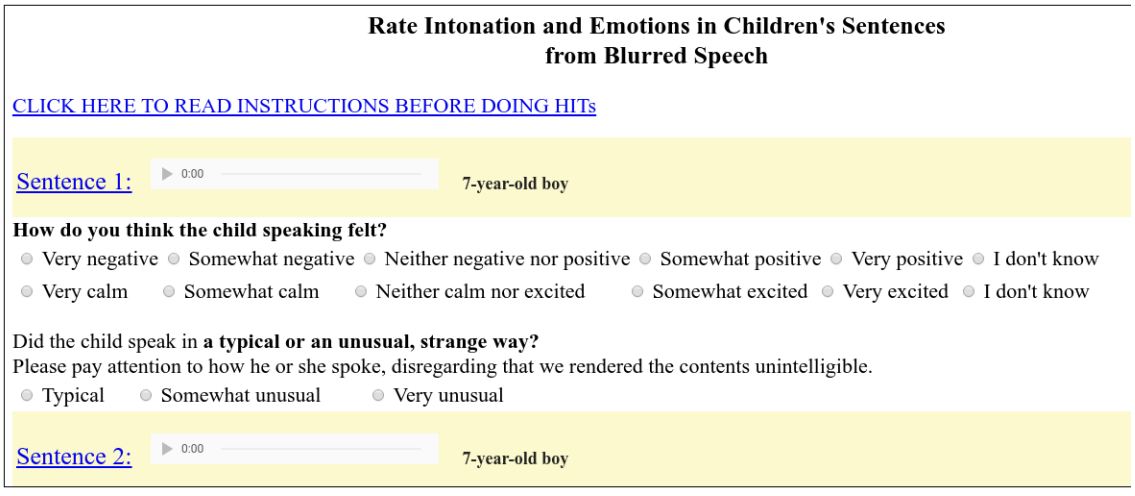


Figure 7.3: Simple (`S`) task interface for the `DELEX` rating task

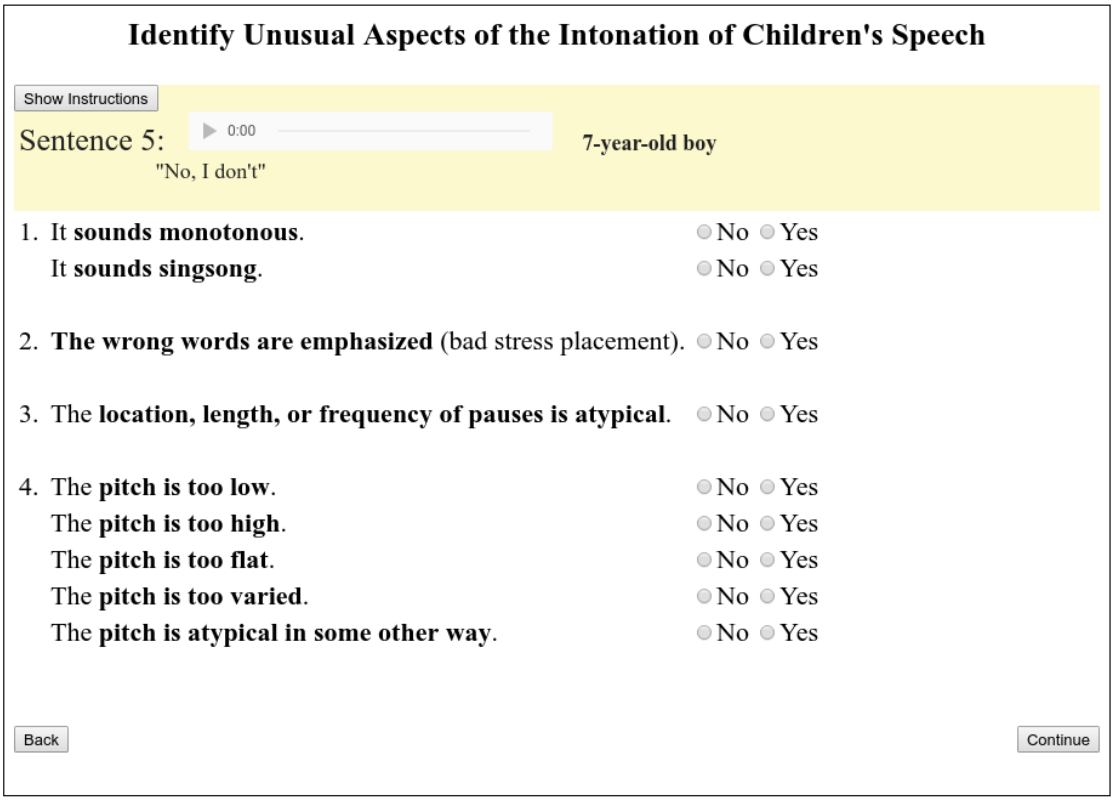


Figure 7.4: Simple (`S`) task interface for the `SPEECH ASPECTS` rating task (questions 1 to 4)

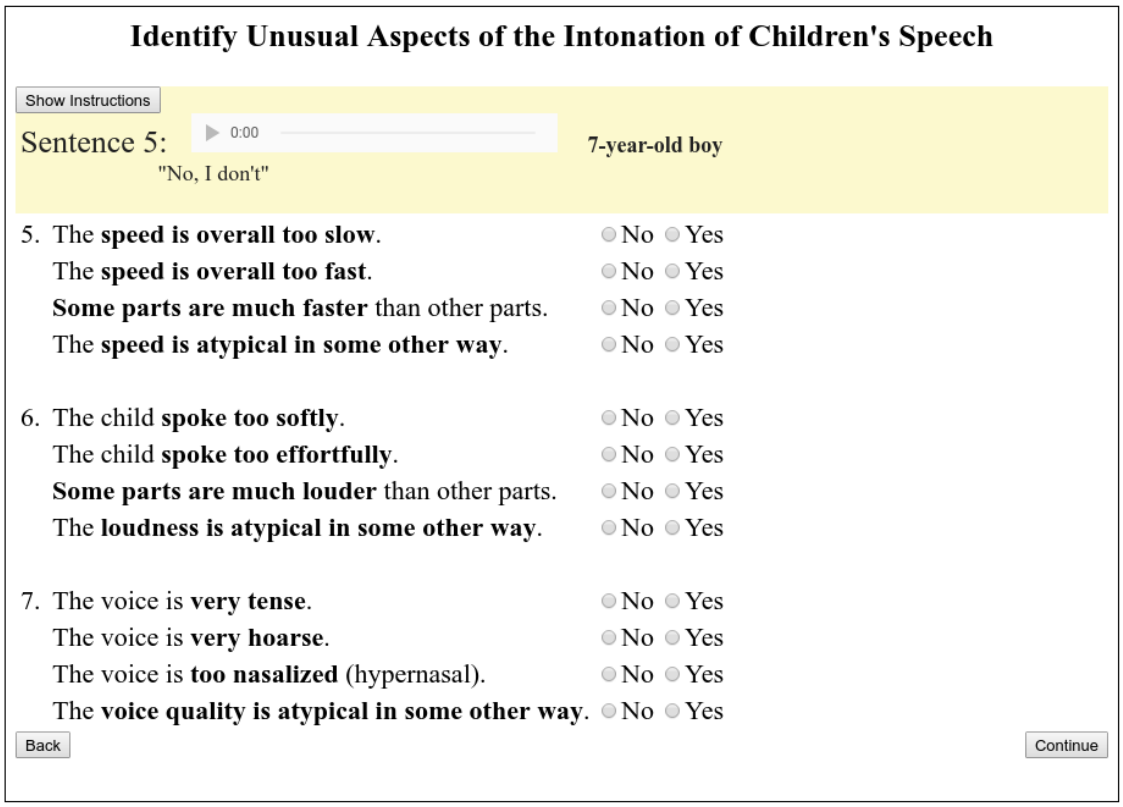


Figure 7.5: Simple (S) task interface for the SPEECH ASPECTS rating task (questions 5 to 7)

Concerns about the simple task interfaces

Colleagues reviewed the tasks and gave feedback to us regarding the instructions and the tasks after the first round of collecting ratings. They noticed various potential shortcomings. We discuss the issues raised and how we addressed them in the points below, resulting in the detailed task interfaces (see the next section).

- People may not be able to provide reliable answers to our question on whether the utterance is typical for a child of a certain age, as they usually do not know much about the stages of child development.

The first version of the tasks (the simple interfaces) displays the age of the child in years (e.g. "4-year-old girl") and asks the above question. We addressed this issue in two ways: First, instead of giving exact ages, we use the "pre-K or kindergarten aged" or "elementary school-aged" expressions, to orient the rater to gross deviations from age expectation, which are presumably easier to judge. Second, we specifically recruited raters who declare that they "have interacted with young children (between the ages of 4 and 8) a lot."

- The words "calm" and "excited" used to explain valence both carry a positive connotation and thus may cue the listeners to interpret emotions as more positive.

Suggested alternatives included "passive" vs. "active" and "low energy" vs. "high energy". Eventually, we used the latter, which has been used in other scientific works as well (see e.g. Jefferies et al., 2008; Tseng et al., 2013).

- The meaning of the words "negative" vs. "positive" for valence may not be obvious to some people; moreover, people can have cultural biases regarding what a negative emotion is.

For lack of better words, we kept those, but made efforts to explain what they mean better through examples, making sure that those examples cover the whole emotional spectrum. Instead of showing two separate rating scales for valence and arousal as before, we started showing the valence–arousal coordinate system to the raters. The coordinate system helps to explain the orthogonality of these concepts, and we illustrated the meaning of specific points in the space by showing some categorical emotion labels. We used a continuous scale, as it possibly gives us more fine-grained information, and is easy to convert to any discrete scale. For the two separate valence and arousal scales, it was one valence and one arousal value (possibly with an "I don't know value" for either or both of them, but the raters hardly ever used those). For the coordinate system, we get two values at once when the rater chooses a point in the coordinate system, and we added a

separate scale for marking the confidence of their rating. Note that the raters need to make two choices in both cases. The new approach seems better in that the raters can indicate being unsure while also giving their best guess for the valence and arousal values instead of choosing "I don't know". They cannot indicate their confidence level separately for the two dimensions though. We can interpret the average value of rater confidence as their self-assessment regarding how good they are at recognizing emotions and the per-utterance average as the difficulty of rating the particular utterance. Finding group differences in the average confidence ratings may be meaningful as well.

- Some sentences may carry multiple emotions, either as an emotion blend or occurring in a sequence, whereas the interface does not allow the raters to indicate that.

Since most utterances are short, we do not expect multiple emotions to occur one after the other frequently to an extent detectable by the raters, and we assume that real emotion blends are also rare. So we decided to keep our approach unchanged, which is to ask for the dominant emotion.

- We considered using the Self-Assessment Manikins (Bradley and Lang, 1994) as an aid to the raters (see Figure 7.6).

We abandoned this idea. The reason was not that ours is not a self-assessment task, as these images might still help in understanding the emotional concepts. However, feedback from colleagues was mixed: Some said that the images were meaningful to them and helped them understand the tasks better, others said that either they did not add anything or were outright confusing (e.g. the seemingly exploding man for high arousal may carry negative connotations, even though it should be independent of valence). Moreover, it does not seem necessary to indicate these concepts each time the rater makes a choice.

- We considered using emoticons for representing the emotions.

They may make it easier for people to relate concepts they use every day to the task at hand. However, it may be impossible to find ones that are equally meaningful to everyone. So eventually we abandoned that idea as well.

- The granularity of the rating scales was not fine enough.

Some questions on atypicality were yes/no questions, still others had three answer options. We switched to using the latter everywhere consistently, thereby increasing the granularity.

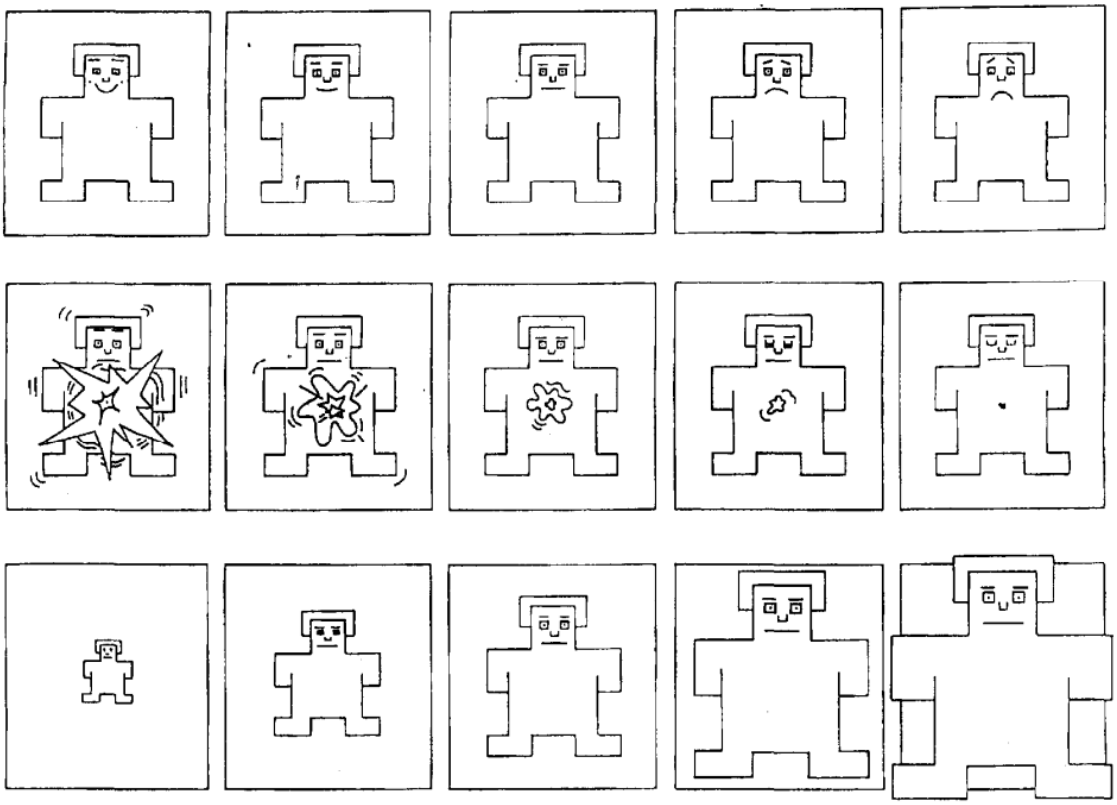


Figure 7.6: The Self-Assessment Manikins (Bradley and Lang, 1994). The rows contain representations for different levels of valence, arousal, and control, respectively.

- The raters only needed to say if the sentences had any unusual words, not which ones those were.

The idea naturally presented itself that we could ask them to mark the specific words that seem unusual to them. The raters must have specific words in mind when answering, and asking about the identity of those should encourage them to give thoughtful answers. Moreover, this gives us the opportunity to see more clearly if they agree with each other and enables us to know not just whether there are such words (a binary answer), but also how many such words there are. It also results in a data set of unusual words, which is of potential value in itself.

- The instructions needed to be made clearer and more concise.

We clarified the explanations based on the feedback received and the included new examples. For example, we now point out that discussing a topic that is unusual for a child (e.g. too advanced) is one of the reasons for marking it as having an "unusual meaning". We also explain prosody by including a list of its constituents ("intonation, rhythm, loudness, speed, voice quality"). We give examples for most other questions as well, including the one asking about the mismatch between prosody and contents.

Detailed task interfaces (`D`)

We updated the task interfaces based on the considerations discussed above. The complete instructions and questions for the detailed task interfaces are available in Appendix A.2. Figures 7.7 through 7.9 show some screenshots of the rating interfaces.

Subsequently, we conducted a usability test: We asked a set of raters to rate some utterances and give feedback on the updated interfaces regarding the ease of understanding the instructions, the wording, handling the interface, the difficulty and likability of the tasks, and anything else they thought was worth mentioning. They confirmed that the instructions were clear and the task interfaces were easy to use. Some of them pointed out that they rarely need to indicate the presence of unusual words or atypical prosody, or that the stimuli are sometimes too short for them to be able to provide a reliable answer. These are inherent properties of the stimuli we worked with, however, there were many long utterances and many with atypical content or prosody as well. A few of them wished that we would show them examples for typical and atypical prosody. We did not address this, as showing a few examples that we consider characteristic could have biased their answers toward our expectations. Instead, we included a request to the raters that they should undertake the task only if they have had ample experience with children in the age range represented in the corpus.

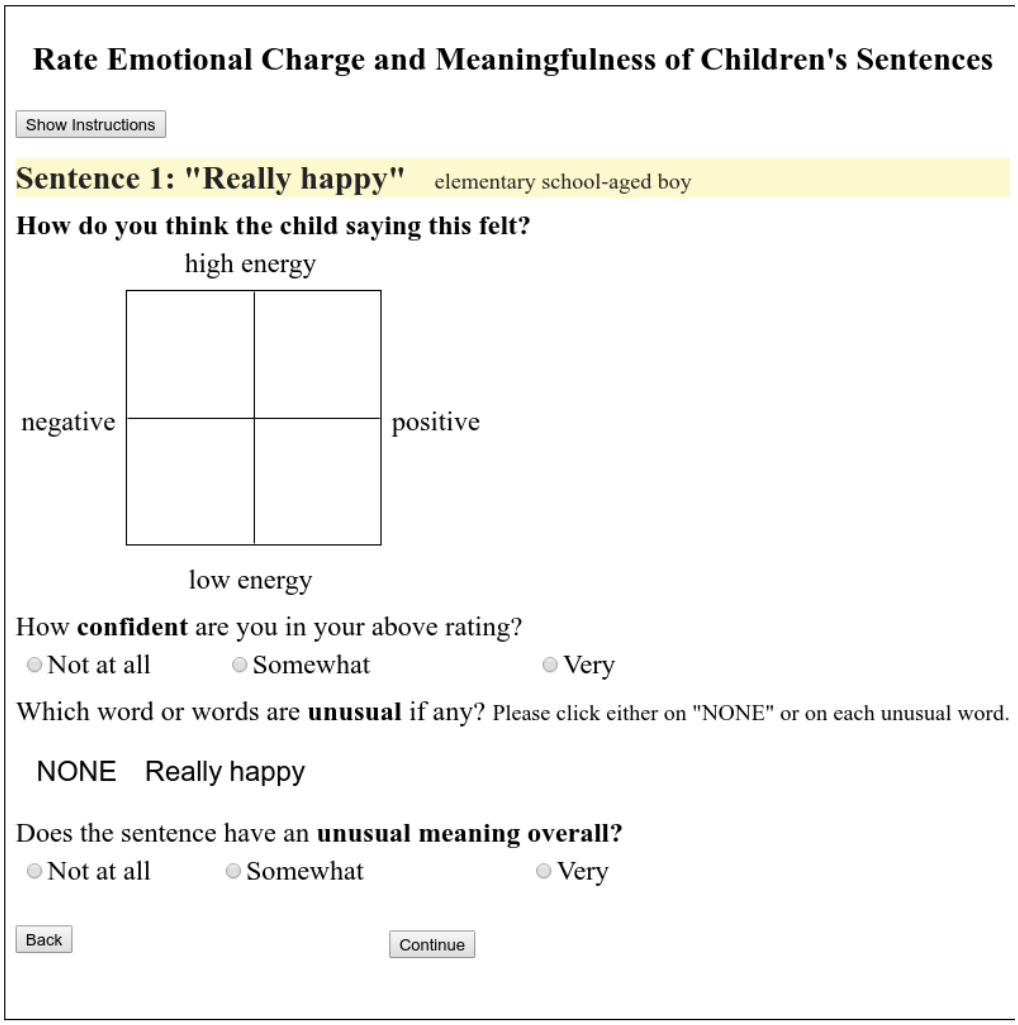


Figure 7.7: Detailed (`D`) task interface for the `TEXT` rating task

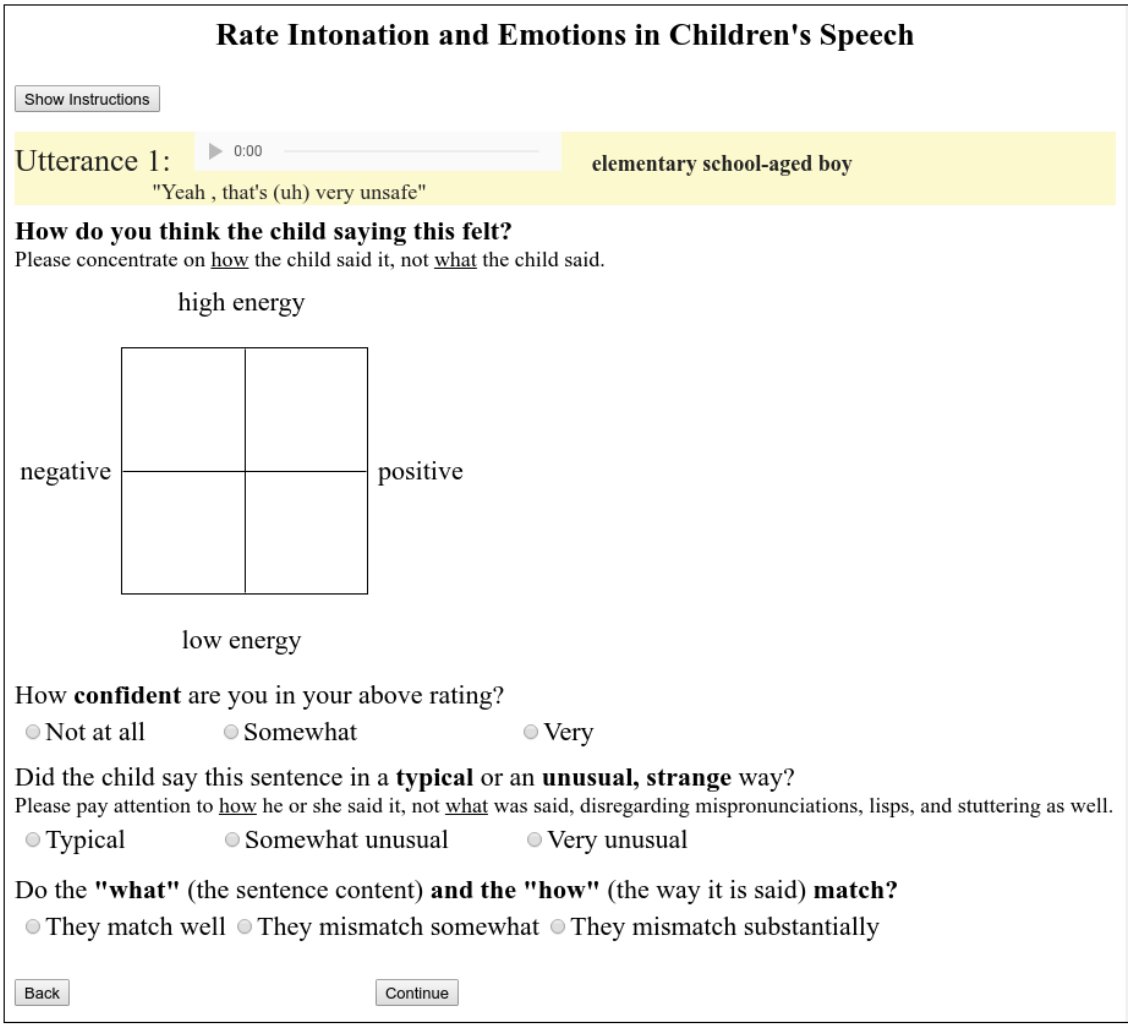


Figure 7.8: Detailed (`D`) task interface for the `SPEECH` rating task

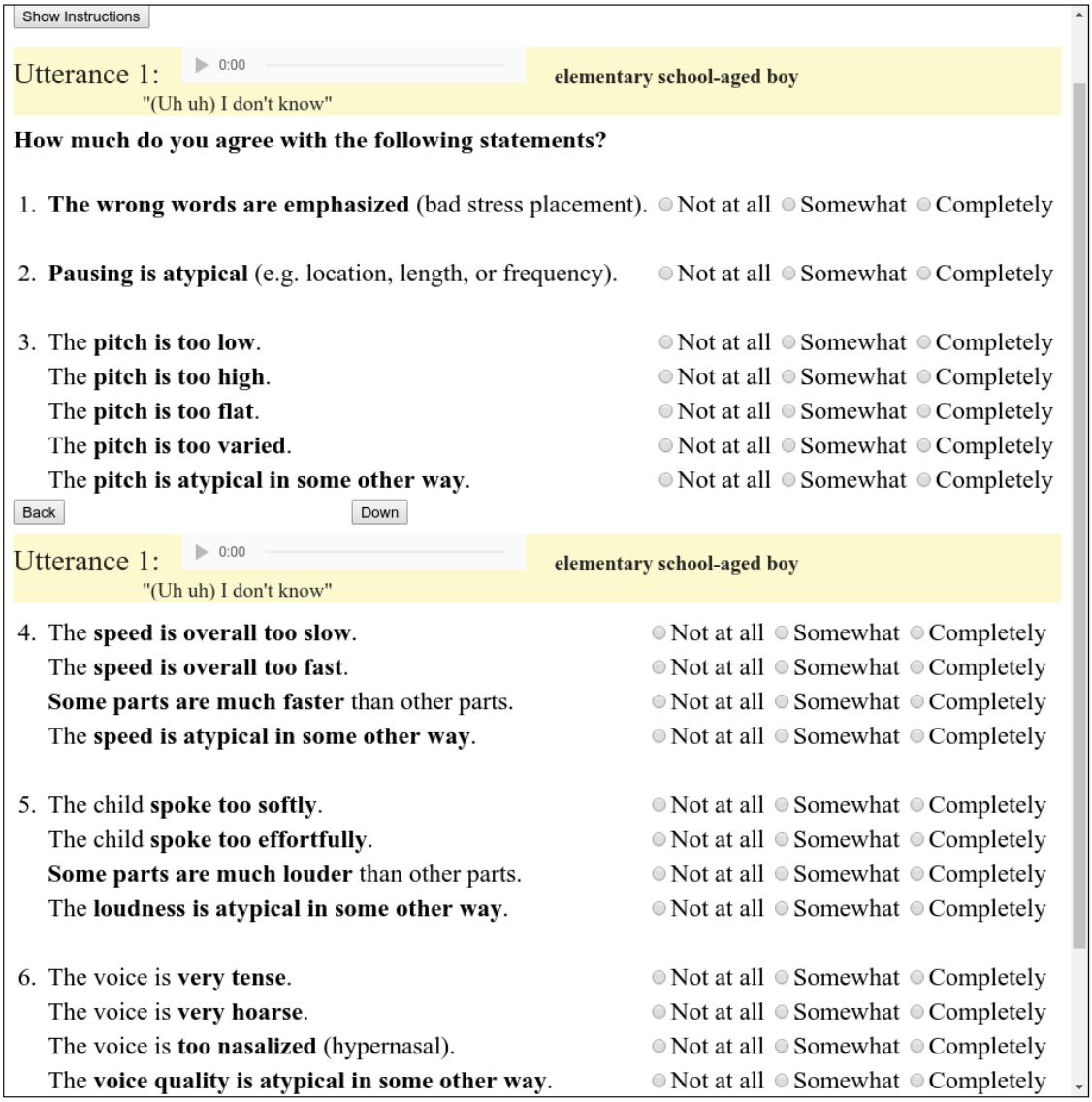


Figure 7.9: Detailed (`D`) task interface for the `SPEECH ASPECTS` rating task

7.2.4 Stimulus sets

The ideal data set for our questions would contain a large number of utterances for a large number of well-characterized subjects from a relatively narrow age-range and with diverse diagnostic conditions, with each utterance being rated on various scales by multiple raters who strive to give thoughtful answers and are not exhausted in the process. In practice, because of financial and time constraints, we can choose a limited number of utterances and have them rated by raters who may not pay full attention to the task. However, we do have relatively large and well-characterized speech corpora.

Below we describe how we selected two sets of utterances, as outlined in our study design. We list further criteria and some details of the technical implementation separately for the two stimulus sets. But first, let us see some general considerations for choosing the utterances.

For both stimulus sets, we excluded echolalic and palilalic utterances based on the SALT markup: echolalia (the spontaneous repetition of another person's speech, in this case the examiner's), as it do not reflect the speech production of the child, and palilalia (the involuntary repetition of one's own words), as it is by definition atypical. Moreover, we did not choose utterances containing grammatical errors or sound effects. We also excluded utterances containing incomprehensible words (as determined by the transcribers) because we cannot fulfill one of our goals, which is to control for textual content. We excluded utterances that overlap with the examiner's speech as well, because we want to be able to work with automatically extracted acoustic features as well, which cannot yet be calculated reliably for such speech.

7.2.5 Stimulus set Matched on Expected Prosody (`MEP`)

Motivation Previous studies usually selected short continuous speech segments of similar size for the subjects, irrespective of content (see e.g. Peppé, 2006; Sharda et al., 2010; Bonneh et al., 2011). Matching the samples only on the length of the speech segment is presumably sufficient for human raters, with more reliable ratings for longer speech segments. However, it may not be optimal for automated methods, because content, as well as the emotional state of the speaker, obviously has an effect on prosody.

Our goal was to compare diagnostic groups pairs through utterance pairs matched on *expected prosody* (see more below). We wanted to see if the utterances for matched diagnostic groups have significantly different prosody when they are as similar as possible in content, including its emotional aspects. We can say that for such utterances, the expected prosody is similar, that is the prosody that one would expect just looking at the transcript for the utterance, so any differences

in prosody are most likely due to prosodic differences between the groups.

Approach To make the expected prosody equivalent for two non-identical utterances, utterance selection would need to match all aspects of prosody (lexical, syntactic, affective, pragmatic, and indexical ones; see Section 2.2), which is clearly not possible for the indexical property but can be approximated for the rest. We take care of matching the grammatical function by choosing utterances with identical utterance type and part-of-speech (POS) sequence. We deal with the lexical similarity as well by selecting utterances with as many identical words as possible. (We did not have information about the semantics or exact pronunciation of the words, so we could not distinguish homographs, such as the noun and verb senses of the word "record".) Thirdly, we choose utterances whose emotional content is similar, as determined by raters based on the text only. Prosody has multiple pragmatic aspects, including "turn-taking control, negotiating agreement, signaling recognition and comprehension, managing interpersonal relations such as control and affiliation" (Ward, 2004). Some of these are captured by the stress pattern, so we make efforts to match the utterances on the expected stress pattern, as predicted by the BioSpeech text-to-speech synthesizer. Matching the pragmatic aspect is not complete: We do not show raters the context, as that would require a very long and complex rating procedure. Notwithstanding, the utterance pairs chosen this way should have well-matched expected grammatical, lexical, affective, and pragmatic.

We used a multi-step process to find the utterance pairs. We first selected utterance pairs whose textual content is similar, as described above, aiming to find 20 utterances per subject when available (possibly more if necessary). We then collected emotional ratings based on the content only using the `TEXT` task, and chose a subset of the utterance pairs that are most similar on these ratings, not just on the emotion but the unusual content ratings as well, aiming to keep 10 utterances per subject. We collected ratings for the prosody of this narrower set in the `SPEECH` and `DELEX` tasks. As seen in Section 7.2.3 earlier, one of the questions was whether the prosody seemed atypical. In the last step, we asked raters to identify which aspects of prosody are atypical for only these atypical-sounding utterances. An illustration is given in Figure 7.10.

For example, our algorithm chose this utterance pair for the subject OGI-092 (ALN) and OGI-161 (TD):

```
"I don't have a lot of friends ."
"It doesn't have a lot of trees ."
```

These utterances were rated in the `TEXT` task, but were not selected to be rated in the `SPEECH` task, presumably because the first one was rated as having a much more negative valence value.

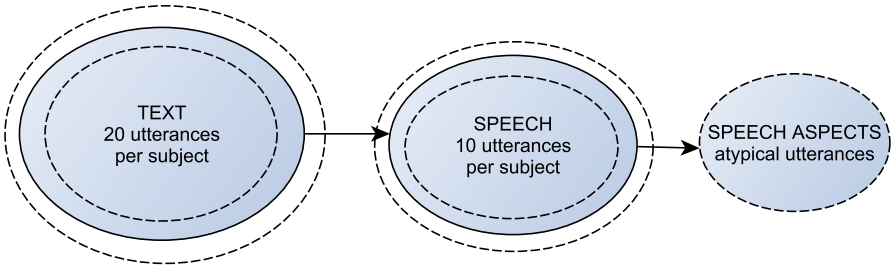


Figure 7.10: The `MEP` stimulus set: Illustration for deriving the utterance sets for the tasks. 20 utterances per subject are chosen for the `TEXT` task and 10 of these for the `SPEECH` task, or possibly somewhat more or less depending on availability of pairs. The utterances deemed atypical are rated in the `SPEECH ASPECTS` task.

Comparison groups We selected utterance pairs for subjects for the matched group pairs ALN–TD and ALI–SLI. These are group-pairs matched independently of each other (see Section 4.2.4) for historical reasons: When we collected the ratings, not as many subject were available, and some of the cognitive scores were not yet finalized; besides our work on matching was not yet done, so we matched the groups using less sophisticated methods that produced a different result.

Both the TD–ALN and SLI–ALI pairs are matched on age at the significance level of $p = 0.4$. The number of subjects in the groups are: 41 TD, 24 ALN, 19 SLI, and 22 ALI subjects.

Utterance candidates We used the manually produced SALT annotations (see Section 4.2.5) to identify the non-overlapping, error-free full utterances with at least three words and no incomprehensible words, no sound effects, and no mazes (disfluencies that do not contribute to the meaning of the utterance, including the pause fillers “um” and “uh”, incomplete words, and revised speech). We narrowed down the set of utterances to those whose POS sequence appears in both matched diagnostic groups (i.e., both TD and ALN, or SLI and ALI). We found 12,069 such utterances in the CSLU ERPA ADOS Corpus, 31–33 utterances per subject on average.

We removed the utterance-final punctuation marks from the transcriptions. The reason is that the punctuation marks reflect the prosodic judgment of the transcribers; for example, an exclamation mark suggests an excited emotional state. We expect the raters to judge the emotional content of the utterances among other things, and we did not want to affect their judgment by these.

Choosing utterance-pairs for the `TEXT` task We created all possible utterance pairs, taking one utterance from each group, both having the same POS sequence. We assigned a weight to each

pair, with a higher weight for more similar content and longer utterance length. Two utterances got the highest weight if their words and expected stress placements were identical, and got decreasing weights with increasing number of differences. We used a greedy algorithm to select 20 utterances per subject when available (possibly more for some speakers, as dictated by the criterion to have matched utterance pairs), adding the utterance pair with the highest weight in each step. For an illustration, see Figure 7.11.

We calculated the weight for the utterance-pair consisting of utterances u_{DX1} and u_{DX2} for the diagnostic group pair as, for speakers S_{DX1} and S_{DX2}:

$$w_{u_{DX1},u_{DX2}} = 1 + C_t \cdot \frac{|\text{identical tokens}|}{|\text{tokens}|} + C_s \cdot \frac{|\text{identical stress}|}{|\text{tokens}|} + C_l \cdot |\text{tokens}| + C_n \cdot (v_{S_{DX1}} + v_{S_{DX2}}).$$

For identical tokens, we count the punctuation marks as well; for identical stress, only the tokens that can carry stress (e.g. "you're").

The last term helped ensure that the number of utterances per speaker would remain close to the expected number of utterances and that it would start matching the more constrained cases. Function v is the weight for a speaker S, which is higher for speakers with less available utterances and more utterances to choose from:

$$v_S = \frac{|\text{utterances to choose}|}{|\text{available utterances}|}.$$

We performed a grid-search for the optimal settings of the parameters (C_t, C_s, C_l, C_n) maximizing metrics that characterize the goodness of the chosen set of utterance pairs. The metrics we used were: the ratio of completely identical stress pairs, the proportion of utterances relative to how many are required (20 per speaker) to the total number of utterances, and the average number of words per utterance.

Subselecting utterance-pairs for the `SPEECH` task We selected a subset of the utterance pairs to acquire prosodic ratings for the audio of this smaller set only, taking into account the ratings we received for the utterance text. We kept 10 of the 20 utterances per speaker (sometimes more if necessary to keep speaker pairs) that maximized the similarity between the affect ratings. We also encouraged the algorithm to choose pairs with similar speaking rate (measured as the number of phonemes per second) and discouraged it from choosing very short utterances. We did this by assigning a score to each utterance pair, and choosing the ones with the highest scores. We calculated the score for each utterance pair as

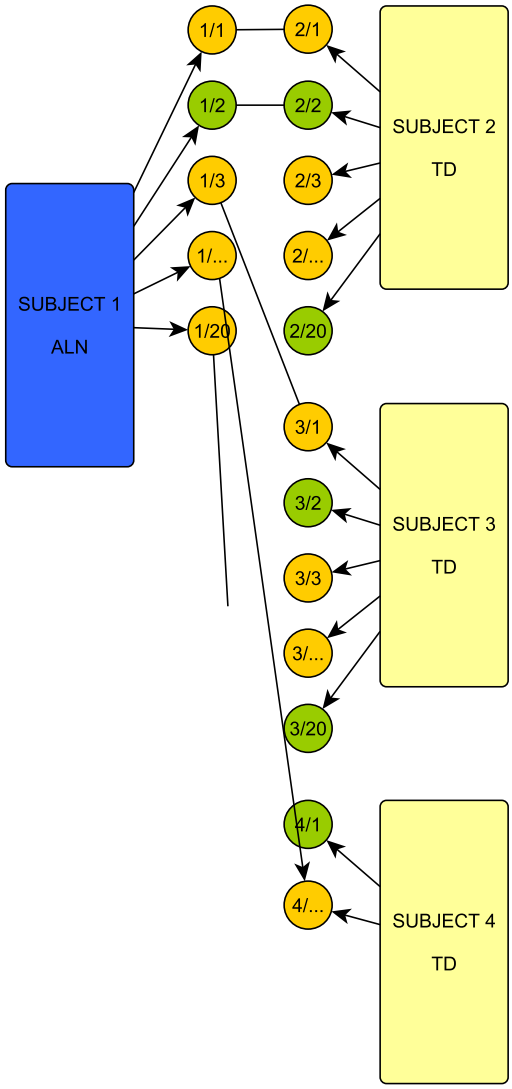


Figure 7.11: The MEP stimulus set: Utterance set sizes per task for a subject. The circles represent the utterances chosen for each subject. The pairs are utterances whose text is identical or very similar. All are rated in the TEXT task and the green colored ones represents the ones that are most similar on TEXT ratings and are thus selected to be rated in the SPEECH task.

$$|\text{valence}_1 - \text{valence}_2| \cdot C_v + |\text{arousal}_1 - \text{arousal}_2| \cdot C_a +$$
$$|\text{stress differences}| \cdot C_d + \left(\text{phonemes per sec}_1{}^2 + \text{phonemes per sec}_2{}^2\right) \cdot C_p.$$

The constants C_v, C_a, and C_p served to limit the maximum value of each term to 1.

7.2.6 Stimulus set that is Maximally Individually Diverse (`MID`)

Motivation The second stimulus served to cover a broad range of different utterance types and lengths, including various roles the utterance plays in the conversation (such as acknowledgment, agreement, disagreement) and having as varied prosodic features as possible. Our goal was to gauge the whole prosodic repertoire of our subjects.

Approach The utterances were chosen for each subject independently (unlike for the `MEP` set). We came up with 33 utterance bins to work with, and chose one utterance for each bin for each subject when available (but not every subject had utterances in each bin). We also chose some more utterances with as varied prosodic features as possible, filling up the total number to 40 utterances per subject. See an illustration of the utterance sets in Figure 7.12. We conserved the code for choosing the utterances in the R package `GUttChooser` (see B.1.8).

Utterance candidates We excluded utterances containing errors, sound effects (e.g. "grr"), non-speech sounds (e.g. lip smack), incomprehensible words, and those overlapping the examiner's speech, but included those with mazes.

Utterance bins When choosing utterances for each bin, we chose ones that are representative of the subject's utterances that fit that bin, preferring ones that fit only that particular bin and not multiple ones at the same time.

We defined the concept of an utterance being the most representative for the particular subject the following way: We used as features the number of words, incomplete words, mazes, and pause fillers, and transferred them to the percentile value within the bin. We defined a metric for calculating the distance between two sets of features: the RMS for the difference of the two feature vectors. For each bin, we determined the mode of our features for the utterances that fit that bin. We designated one utterance as the most representative one in the bin if the distance of its features from the feature modes was the smallest.

The utterance bins were the following:

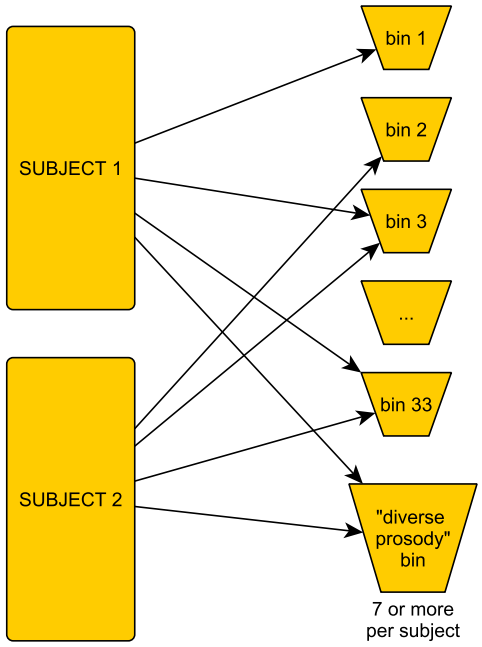


Figure 7.12: The `MID` stimulus set: Illustration for deriving the utterance sets for the subjects. For each subject, we chose one utterance per bin when available and as many prosodically diverse utterances as necessary to have 40 utterances per subject.

1. positive acknowledgment: a “Yeah”, “Yes”, or “Yep” answer (in this order of preference) to the examiner’s yes/no question
2. negative acknowledgment: a “No” or “Nope” answer (in this order of preference) to the examiner’s yes/no question
3. agreement: a “Yeah”, “Yes”, or “Yep” answer (in this order of preference) to the examiner’s statement
4. disagreement: a “No” or “Nope” answer (in this order of preference) to the examiner’s statement
5. sentence starting with a discourse marker: “And ...”, “But ...”, or “Well ...” (in this order of preference); preferably turn-initial
6. an “uh” pause filler in initial position
7. an “um” pause filler in initial position
8. an “uh” pause filler in medial position
9. an “um” pause filler in medial position 10-33. all possible combinations of: 4 word number ranges (1; 2–4; 5–6; 7–10) × 3 sentence types (S; YN; WH) × 2 maze conditions (present; missing).

The above list requires some more explanation. When there was more than one candidate word, we preferred the ones that were present for more subjects in at least one utterance, to make the utterances in the bin as uniform as possible. The boundaries of the word number ranges approximately correspond to the 25th, 50th, and 75th percentiles of the number of words within the utterances.

We determined the sentence type automatically using heuristic rules based on surface features and the output of the Stanford parser, as described in Section 4.2.7. After we ran the sentence selection algorithm, we reviewed all utterances and corrected this label when necessary. Since the sentence type is one of the features that are taken into account during the selection process, as long as any of the labels were wrong, we reran the sentence selection algorithm and reviewed the labels again. Obviously, as long as there are incorrect sentence type labels in the corpus, some sentences may not be considered for their actual target group. However, at least this iterative process resulted in a set of utterances that truly meet our criteria.

Prosodically diverse utterances We chose utterances with as varied prosody as possible by choosing additional utterances that are "farthest" from already chosen ones acoustically. We chose the same distance measure as described above, and as features, we used statistical features of the fundamental frequency and intensity curves for each utterance. The statistical features were: 90% quantile, median, inter-quartile range (IQR), IQR / median (for a robust measure of the coefficient of variation), robust skewness (Kim and White, 2004, SK_2), and robust excess kurtosis (Kim and White, 2004, KR_2). As long as we had less than 40 utterances for a subject, we added one more for which the RMS of its distances from the already chosen utterances was maximal.

7.2.7 Audio normalization

The speech volume of the utterances originally varied greatly from subject to subject and also within subject, due to varying settings of the recording devices and the child moving around in the room occasionally. It was necessary to normalize the volume of the recordings so that the listeners would not need to keep adjusting the volume of their playing devices for each utterance. We normalized the audio for the utterances by bringing the 90th percentile of the RMS energy from all speech segments to the same level for all recordings. The corpus contains multiple recordings for most speakers: two channels from two external microphones, and occasionally a recording from the camera microphone. The raters got to listen to the normalized version of the loudest unsaturated speech segment that is available for the utterance.

7.2.8 Rating collection

In an ideal situation, a large number of highly skilled people would attentively read the task instructions, listen to all the speech stimuli, and then carefully answer our questions. In actuality, partly due to time and financial constraints, we can only get a few people of varying skill levels to listen to some (but not all) of the stimuli, and they get tired while doing the tasks. We need to find ways to assess the quality of the data we managed to collect in the circumstances.

We published the rating tasks and recruited raters through the Amazon Mechanical Turk (AMT) crowdsourcing platform. This has the benefit that, once the stimuli and the task interfaces are ready, one can recruit workers for a particular task effortlessly from all over the United States (and indeed the whole world), and quickly collect a large amount of data for a reasonable price. The drawback is that one cannot know the workers personally, which makes it harder to assess how competent they are and whether they are determined to do their job thoughtfully, as some people submit random answers (spam) to reduce the time spent and to maximize their income.

This can be an extra source of noise beyond what comes from task difficulty and potentially the task being underspecified. Therefore we made careful preparations and monitored and evaluated their work to ensure that the results are reliable. Below we describe the process, the safeguards we put in place, and how we evaluated the data quality.

Publishing the tasks We submitted the utterances to be rated to the AMT system in multiple batches. Each batch contained multiple so-called Human Intelligence Tasks (HITs): a rating task for a certain number of utterances, denoted hereafter by `UPH` = number of utterances per HIT. We randomly ordered the utterances, but put utterances for children with the same age and gender into every HIT when possible, to ease the cognitive load on the raters, as we expected them to judge the typicality of the utterances compared to age and gender expectations. (We did in fact get feedback from a rater saying that they valued this decision.) We asked a certain number raters, denoted by `APH` = number of assignments per HIT, to do each HIT. This enabled us to calculate inter-rater agreement, to estimate rater competencies, and to be able to come up with more reliable ratings than any of the raters individually would be able to attain, similarly to the approach used by van Santen et al. (2009). The total number of assignments is thus the number of utterances divided by `UPH` times `APH`.

We made sure that the task implementations worked reliably. We tested them in multiple browsers (Firefox, Konqueror, Google Chrome). We used JavaScript to implement the page logic. If JavaScript was not available, we did not display the task, only showed a message that asked the potential employee not to take the task. We did not let the raters submit their work before they answered every question, and if some answers were missing, we highlighted the missing information.

Selecting raters We recruited raters who seemed well-suited to do the tasks. We only allowed workers from the United States, to ensure that they are native speakers of (or at least accustomed to) US English, who have at least 98% approval rate, and have completed at least 10,000 tasks. Since we collected the ratings in multiple batches, we were able to glean information about their reliability. For later batches, we excluded a few raters who submitted spam earlier, and specifically invited the raters who seemed to be the most reliable ones, while not excluding others either from doing the tasks.

The raters were blind to the diagnosis of the children and to the purpose of the study. We asked them to read instructions before doing the tasks that explained to them what they needed to pay attention to. We also required them to do a qualification task to check their understanding of the instructions.

Monitoring raters We recorded all page actions of the raters, and used that to confirm if they were reliable workers. We received their time-stamped list of actions (when they listened to each speech recording, when they gave their answers or pressed any button) with each submission, together with their ratings. We analyzed these to see whether they listened to the speech samples that they rated, how much they thought after listening, how much time they spent on rating each HIT in total, among other things. In a few cases, raters did not listen to some speech samples (perhaps accidentally), or they spent hardly any time on answering our questions. When this happened, we did not use their answers, prevented them from answering our future tasks, and uploaded these stimuli to the crowdsourcing system again to get new answers for them. For the raters whose responses we did not reject entirely, we still estimated their competence. The decisive factor in this case was not the time spent on the tasks or other similar metrics, but how well they agreed with each other, as described in Section 7.2.9.

Getting additional ratings For one of the data collection batches, we increased the reliability of the aggregated ratings by getting more ratings for the utterances with relatively large standard error. More specifically, when the standard error for any rating or for the average of all ratings was above the 85th percentile we asked for five more sets of ratings from new raters for the `MID-S` set (see Table 7.1 later).

7.2.9 Aggregating ratings

Source of variability in the ratings We collected multiple ratings for each utterance to be able to come up with a reliable aggregate score. The variability of the answers can come from multiple sources: ambiguity in the instructions, multiple possible interpretations for the rated item, inter-rater variability (differences in personality or competence between the raters), and even within-rater variability (the same person may judge the same situation differently depending on mood, energy level, external influences, etc.) We strived to make the instructions clear and unambiguous. We made it possible to indicate the level of certainty for the answers where it seemed warranted, namely for the emotions, as it must be often hard to decide that (especially when the utterance is short or only the text is available). Consequently, our expectation is that most of the variability comes from differences in the personality of the raters or their competence levels. If indeed it is only a difference in their personality, then each rating should be taken into account with the same weight. If the more important factor is their competence level, evidenced by a rater disagreeing with other raters even on obvious cases, then giving more or less weight to some workers seems justified. In our case, we expect rater competence to vary besides having

items that are hard to score.

Levels of measurement For most questions, the raters were to choose from a discreet number of choices (2, 3, or 5) on a Likert-like scale. Our scales are symmetric and we intended the choices to be equidistant. It is common practice to interpret such data as interval-level data rather than merely ordinal. The emotion ratings collected using the detailed task interface were on a continuous scale, which we can assume is interval-level.

Calculating aggregated ratings Assuming that the ratings have an interval nature, we can combine them using a weighted mean and potentially a bias-term. For R raters with rating scores $s_r, r = 1 \ \ldots \ R$, weights w_r and biases b_r, the aggregated score is:

$$\hat{s} = \frac{\sum_{r=1}^{R} (s_r - b_r) \cdot w_r}{\sum_{r=1}^{R} w_r}$$

The rater competence reflects the skill and reliability of the rater, and the bias any systematic deviation in the expected value of his or her ratings.

There are multiple ways we can set the weights and biases:

1. If for each $r = 1 \ \ldots \ R$, we use $b = 0$ and $w = 1$ then this is simply the average.
2. If $b = \frac{1}{R} \sum s_r$ and $w = \mathrm{SD}(s_r)$ for $r = 1 \ \ldots \ R$, then it becomes the mean of the z-normalized scores.
3. If b_r and w_r depend on the identity of the rater, then it can take rater competence and bias into account. Researchers have used all of these approaches (see e.g. van Santen et al., 2009; Ipeirotis et al., 2010; Bone et al., 2015).

In theory it is possible that the rating for some utterances is multi-modal, in which case a weighted mean is far from any value chosen by the raters. But one would need a large number of ratings to decide if this is the case, which is not available to us. Moreover, this is presumably a rare case, so we ignore this possibility in this study.

If we consider the ratings to be on an ordinal scale, then other methods are more appropriate. For example, it can be the majority vote, or the median. Researchers have developed more sophisticated approaches as well that take into account rater competence, for which we show some examples below.

Estimating rater competence and bias Several studies investigated how to estimate rater competence and aggregating multiple ratings to get a more reliable estimate of the target measure (Sheng et al., 2008; Hsueh et al., 2009; Whitehill et al., 2009; Welinder et al., 2010; Ipeirotis et al., 2010; Ipeirotis, 2011; Lin et al., 2012; Liu and Wang, 2012; Ipeirotis et al., 2013; Wang et al., 2013). We were able to attain and use the software for only one of them, namely Get Another Label (Ipeirotis et al., 2010), but eventually decided not to use it. Its fundamental idea is to estimate rater properties and item difficulty using an Expectation-Maximization framework. There are two versions available: one for combining nominal labels, which seems to work robustly, and one for combining continuous scores, which we were not able to use. Regarding the version for nominal labels, first of all, it does not exactly match our needs as our ratings should rather be considered ordinal or interval variables, second, the result depended largely on the number of iterations for our data. The rater competence estimates converged either to zero or to one as we increased the number of iterations, making it practically impossible to determine an optimal solution. Therefore we implemented our own approach.

We can picture the collected ratings as a sparse matrix, with the utterances corresponding to the rows and the raters corresponding to the columns. Each rater has at least `UPH` items in common with `APH` other raters. For such a block, we can use as rater competencies the first eigenvector from the Principal Components Analysis (PCA) applied to the covariance matrix of the rater (van Santen et al., 2009, p.1086, Section 3). But some raters do more than `UPH` items, even if those items are (more often than not) rated by a diverse set of other raters. We created an approach that took advantage of this fact.

We used an iterative approach to estimate rater competence and bias. The procedure starts with uniform competence and bias estimates for all raters, then iteratively updates these estimates based on the relationship between the rater scores and the aggregate score calculated using the current competence and bias estimates. We can see the outline of the algorithm in Figure 7.13. We experimented with two approaches for updating the competence value estimate: The product-moment correlation of the aggregate score and the rater score, which is included in the figure, and the one described by Warfield et al. (2006).

The reliability of these methods might be checked by using gold standard scores, but we do not have that. We do have external evidence about the competence of some of the raters however: We know that the ones who did not listen to utterances and those who spent hardly any time on the tasks cannot have done a thoughtful work. We did an informal assessment by checking the competence values calculated for these “spammers”. We found that the correlation-based method gave competence values close to zero for such raters, whereas the values from our implementation

1. Initialize $b_r = 0$ and $w_r = 1$ for $r = 1 \ \ldots \ R$
2. For each utterance $u, u = 1 \ \ldots \ U$, calculate aggregated score $\hat{s}_u$ from $s_{i,r}$, w_r, and b_r, where the set of utterances are $\mathbf{u}$, those rated by rater r are $\mathbf{u}_r$, the ratings scores are $s_{i,r}, i \in \mathbf{u}_r$.
3. Re-estimate rater biases:

$$b_r = \frac{1}{|\mathbf{u}_r|} \sum_{i \in \mathbf{u}_r} s_{i,r} - \hat{s}_{i_r} \text{ for } r = 1 \ \ldots \ R$$

4. Re-estimate rater competencies:

$$w_r = cor(\hat{\mathbf{s}}_r, \mathbf{s}_r) \text{ for } r = 1 \ \ldots \ R$$

5. Proceed to 2. up to a the maximum number of iterations or until it converges.

Figure 7.13: Correlation-based iterative algorithm for estimating rater bias and competence values

of the approach described by Warfield et al. (2006) did not reflect that outside information. For this reason, we trusted and used the correlation-based method. Note that it is essentially identical to the approach used by van Santen et al. (2009) when used for one full `UPH` times `APH` block of ratings.

7.2.10 Data analysis

Using the steps outlined before, we collected a set of ratings for a sizeable set of utterances from the CSLU ERPA ADOS Corpus. For a summary of the ratings provided for different kinds of utterances, see Figure 7.14 (the underlined names are the identifiers for the particular column). All types of ratings may not be available for all utterances.

We performed planned comparisons for the unusual content and atypical or incongruent prosody ratings. We expected that all of these differ significantly between the ALN and TD groups, and even more so for the HFA–TD comparison (where HFA includes both the ALN and ALI subjects), and that they would not differ for the ALI–SLI comparison. We expected emotional differences between the groups: originally that the `SPEECH` valence rating would be different, but were also going to compare arousal. After an initial data collection, it became apparent that the perceived emotional difference is on the arousal dimension, in agreement with earlier research that analyzed physiological arousal signals (see e.g. O'Haire et al., 2015).

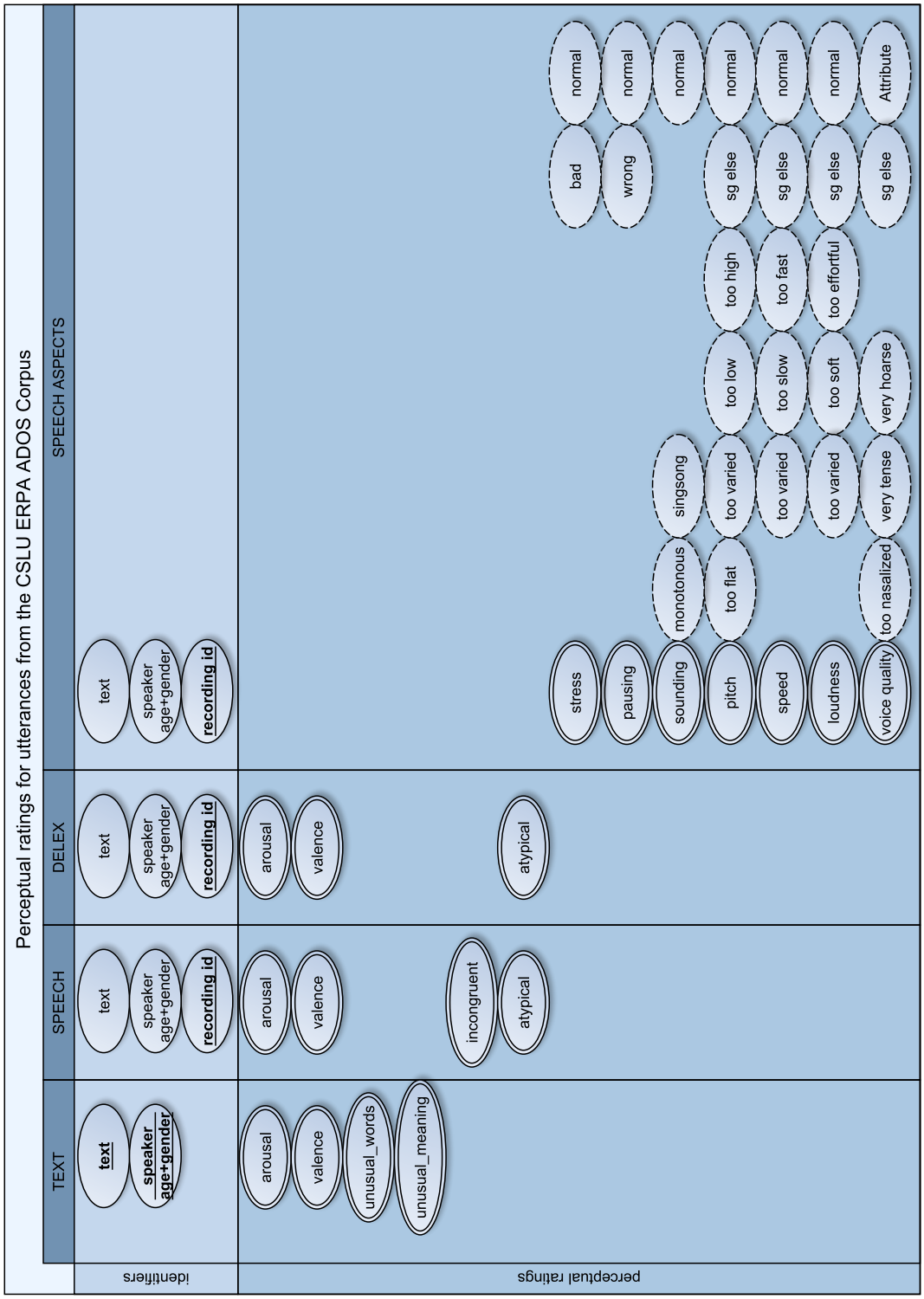


Figure 7.14: Perceptual ratings for utterances from the CSLU ERPA ADOS Corpus

7.3 Analyzing the Perceptual Ratings

7.3.1 Assessing the perceptual ratings data

Below we summarize some important properties of the perceptual ratings we collected for recordings of the CSLU ERPA ADOS Corpus. To see how well-powered the analyses of this data set are, we show the number of rated utterances and of the independent ratings for each one. The level of agreement between raters sheds light on how reliable the data is. We also wanted to see what difference the instructions and the rating scales had on the results.

Number of ratings collected

Table 7.1 contains the number of utterances rated in the first four rows. Each utterance was rated by a pre-determined number of raters, as shown in the "assignments per HIT" row. The `MID-S` data set contains a large number of utterances with six or more ratings because we collected extra ratings for the utterances whose scores had the highest standard errors after getting the first three ones. This strategy worked well, as the overall agreement between the raters is higher for this set than for `MID-D` as we will see later, even though the average number of ratings per utterance is lower. The minimum number of utterances that were all rated by the same set of raters is "utterances per HIT", which number is relevant for calculating inter-rater correlations.

Table 7.1: The number of utterances rated from the CSLU ERPA ADOS Corpus

	`MEP` stimuli, `S` interface	`MID` stimuli, `S` interface	`MID` stimuli, `D` interface
`TEXT`	2168	2446	2935
`SPEECH GENERAL`	1273	4148	4521
`SPEECH ASPECTS`	429	-	419
`DELEX GENERAL`	1273	-	-
assignments per HIT	5 (rarely 10)	3 (up to 9)	10 (rarely 20)
average assignments per HIT	5.07	4.20	10.22
utterances per HIT	10	10	25

Inter-rater agreements

We estimated the average group-wise correlation between sets of raters, similarly to Asgari et al. (2014, Section 2.1), randomly dividing the raters into two groups several hundred times and calculating the correlation between the aggregates for the two groups. This analysis was more complicated in our case because the number of utterances evaluated varied greatly from rater to rater. Our solution was to assign the raters randomly to two groups while trying to keep the number of ratings in the two groups balanced. The algorithm iterated over the raters in a random order, and for each rater, it chose the group that minimized the group difference in the number of ratings, or if that it did not matter, it chose one of the groups randomly. We calculated the correlations between the bias-corrected scores for the two groups 100 times, and averaged the correlations; the result is given in Table 7.2. We calculated the same values by training a linear model on synthetic data to handle the problem of rater biases, which gave very similar results.

Table 7.2: Average correlation between the aggregated perceptual ratings for randomly selected subgroups of raters for the CSLU ERPA ADOS Corpus

	`MEP` stimuli, `S` interface	`MID` stimuli, `S` interface	`MID` stimuli, `D` interface
`TEXT` unusual words present	0.34	0.37	0.56
`TEXT` unusual meaning	0.42	0.46	0.67
`TEXT` arousal	0.33	0.39	0.58
`SPEECH` arousal	0.63	0.68	0.85
`DELEX` arousal	0.42		
`TEXT` valence	0.69	0.70	0.81
`SPEECH` valence	0.53	0.58	0.76
`DELEX` valence	0.24		
`SPEECH` prosody incongruous	0.11	0.19	0.45
`SPEECH` prosody atypical	0.25	0.27	0.49
`DELEX` prosody atypical	0.17		

These correlations vary greatly from data set to data set, and even the highest correlations, those for arousal and valence for the `MID-D` data set, are much smaller than the ones in the CSLU Cross-Model Corpus for a similar task (see Section 4.1). This may be due to a fewer number of raters assigned to each utterance (for the `MEP-S` and `MID-S` columns; for the Asgari et al. (2014)

paper, all speech recordings were rated by the same 11 undergraduate students on 5-point rating scales), shorter utterances, a different speech elicitation method, and less competent raters or spammers playing a role in our case. In the MEP stimulus set, the shortest utterance is made up of three words and the longest of eight, with the average being around four words per utterance. The number of words per utterance is less balanced in the MID stimulus set: over 40% of the utterances consists of only one word. There are also substantially longer utterances, with the average being around three words per utterance; but as we shall see, we control for this utterance length variability in our analyses. Comparing these numbers to the average of over six words per utterance in the CSLU Cross-Model Corpus, it is apparent that the task of our raters must have been harder. Another difference is that a different speech elicitation method was used here (spontaneous speech vs. story retelling in the other case). We can also see that doubling the number of raters increased the agreement substantially (see MEP-S and MID-D with 5 and 10 raters per utterance, respectively). The relationship between the magnitudes for different modalities is similar to that of the Cross-Modal corpus. For example, raters agree more on valence than arousal when rating the text, and more on arousal than valence when rating the delexicalized speech.

In our analyses, we concentrate on the ratings with higher reliability. Whereas the correlations for the first two data sets (MEP-S and MID-S) are between weak and strong, the values for the third data set (MID-D) are moderate to very strong correlations. The prosodic atypicality rating, which is of utmost importance for this study, generally has a low agreement, with a moderate agreement only in the third data set for the SPEECH task. The emotional ratings seem to be quite reliable for all data sets. The aggregate ratings, which are derived from all ratings and not just from half of the raters as here, can be trusted even more.

Effect of the rating interfaces on the ratings

We wanted to see whether the type of interface ("simple" vs. "detailed") and the difference in the associated instructions had an effect on the ratings. For this purpose, we calculated the correlation between the aggregated ratings for the utterances that occurred in several data sets: MEP or MID stimulus sets with the S or D interfaces. Table 7.3 contains the results. The first column is for two data sets collected with the same interface (S), whereas the second and third columns are for differing interfaces (S and D). If one of the data sets was very different from the other two, then the correlation of its ratings with the other two data sets would be much lower then the correlation of the other two data sets with each other. We can see that there is no consistent difference between the columns and that they are very similar to each other across the board. This indicates that the type of rating interface did not have an important effect on the results, other than the data

collected using the D interface being more granular.

Table 7.3: Correlation between the aggregated perceptual ratings for utterances from the CSLU ERPA ADOS Corpus that appear in multiple data sets

Pearson correlations	MEP-S and MID-S	MEP-S and MID-D	MID-S and MID-D
TEXT arousal	0.58	0.61	0.61
SPEECH arousal	0.79	0.79	0.83
TEXT valence	0.85	0.85	0.84
SPEECH valence	0.72	0.70	0.77
unusual words	0.35	0.50	0.46
unusual meaning	0.62	0.81	0.59
prosody atypical	0.54	0.58	0.48
prosody incongruous	0.32	0.28	0.40

7.3.2 Group differences in ratings

Results for the MEP stimulus set

We compared the per-speaker average ratings across the diagnostic group pairs, namely SLI–ALI and TD–ALN, using Monte Carlo tests (see Section 3.4.3) for the following: all variables, only those for the TEXT or the SPEECH ratings, and separately for each variable (see Tables 7.4 and 7.5). No differences were significant for the SLI–ALI comparison. For TD–ALN, all ratings together differentiate between the groups, but not the TEXT ratings, namely unusual words or meaning and the emotions for the text, which was was expected as the MEP stimulus set was matched on the content. We now proceed to look at the various ratings for the other tasks.

The ratings for the SPEECH task were significantly different for MEP-S for the TD–ALN comparison. For the individual variables, arousal differed highly significantly ($p < 0.001$): children with ALN were perceived as having higher arousal while speaking. The p-values for the per-subject means was very similar to this result. None of the other variables (valence, prosodic atypicality, incongruity) were significantly different. Note that we do not do a TD–HFA comparison here, as the HFA (i.e. merged ALN–ALI) group is not matched on utterance content to the TD group.

The group difference was not significant for the DELEX task for MEP for any rating. It seemed that this task was much harder to do reliably than the other ones, not just from the non-significant results, but also seeing that raters were reluctant to undertake this job (the results came in much slower than for other tasks). Because of this experience, we did not expect to get meaningful

results for this task and dropped it from later experiments.

Results for the MID stimulus set

For the MID stimulus set, we worked with groups matched on age only (at $p = 0.2$) similarly to the MEP stimulus set. Again, there were no significant differences for the SLI–ALI comparison. For the TD–ALN and TD–HFA comparisons, all ratings together differentiate the groups, as well as the SPEECH ratings, but not the TEXT ratings. Some TEXT variables seem to differ when looking at them individually, namely the words unusual and meaning unusual ratings, but the differences are not consistent across the data sets. For the SPEECH arousal rating, we found the same trend as what we saw earlier for the MEP-S data set, but the difference is not as pronounced (see Table 7.5).

Table 7.4: SLI–ALI group difference: Monte Carlo p-values for the perceptual ratings. We show the p-values and also mark their significance level (*** 0.001 ** 0.01 * 0.05 + 0.10).

	MEP-S	MID-S	MID-D
all variables	.95	.27	.42
all TEXT variables	.51	.06+	.06+
words unusual	.94	.31	.34
meaning unusual	.75	.43	.59
emotion arousal–TEXT	.26	.24	.09+
emotion valence–TEXT	.13	.08+	.19
all SPEECH variables	.91	.48	.61
emotion arousal–SPEECH	.89	.53	.32
emotion valence–SPEECH	.45	.79	.96
prosody atypical	.64	.60	.70
prosody incongruous	.42	.40	.88

Table 7.5: TD–ALN group difference: Monte Carlo p-values for the perceptual ratings. We show the p-values and also mark their significance level (*** 0.001 ** 0.01 * 0.05 + 0.10).

	MEP-S	MID-S	MID-D
all variables	.001***	.01*	.007**
all TEXT variables	.09+	.34	.11
words unusual	.02*	.34	.05*
meaning unusual	.92	.10	.04*
emotion arousal–TEXT	.32	.87	.24
emotion valence–TEXT	.81	.15	.23
all SPEECH variables	.001**	.008**	.01*
emotion arousal–SPEECH	.001***	.06+	.05*
emotion valence–SPEECH	.57	.28	.66
prosody atypical	.90	.04*	.14
prosody incongruous	.29	.07+	.30

Table 7.6: TD–HFA group difference: Monte Carlo p-values for the perceptual ratings. We show the p-values and also mark their significance level (*** 0.001 ** 0.01 * 0.05 + 0.10).

	MEP-S	MID-S	MID-D
all variables		.005**	.02*
all TEXT variables		.62	.93
words unusual		.70	.66
meaning unusual		.10+	.84
emotion arousal–TEXT		.81	.57
emotion valence–TEXT		.77	.85
all SPEECH variables		.007**	.03*
emotion arousal–SPEECH		.03*	.11
emotion valence–SPEECH		.77	.71
prosody atypical		.007**	.06+
prosody incongruous		.17	.39

7.3.3 Addressing potential pitfalls

Effect of neutral ratings on emotional score differences

One of our findings was the significant difference in the arousal ratings between the ALN and TD groups for the `SPEECH` task ratings collected using the `S` interface, so the question arises if indeed the ALN group received higher ratings or perhaps the difference in the number of neutral ratings distorted the aggregate scores. This latter explanation might be true if for example it is harder to decide the emotions for the children with ASD, and such hard cases might in turn prompt some raters to choose the neutral rating.

To answer this question, we calculated and compared the balance between the positive and negative ratings for the subjects in the ALN–TD comparison groups. Our formula for the balance is

$$B = \frac{N_+ - N_-}{N_+ + N_-}$$

where N_+ stands for the number of positive ratings and N_- for the number of negative ratings. The value of B is between -1 and $+1$ and zero if the number of positive and negative ratings is the same.

We found that the arousal difference between the HFA and TD groups cannot simply be due to a difference in the number of neutral ratings. Comparing the balances between the groups using t-tests showed a similar pattern of differences as seen for the actual ratings: The B value is significantly different between ALN and TD for the `MEP-S` data set, $p < 0.001$.

Effect of activity and sentence type on arousal and valence

Another question related to our result on arousal for the `MEP` stimulus set is whether it just mediates a systematic difference in the ratio of activities or other content features between the diagnostic groups. In other words: If the relative percentages of utterances derived from the respective activities are not the same across diagnostic groups in the stimulus set, then that itself may be a reason for the average arousal difference.

We fit a mixed effect linear model to the data to see if the difference in arousal persist after we control for content features. As fixed effects, we specified the diagnostic group, the activity, and the sentence type, as well as the arousal and valence rating for the `TEXT`, and the utterance duration. As random effects, we had intercepts for the subjects and random slopes for the `TEXT` emotion ratings. This time we did not remove factors with non-significant effects to be able to show their estimated effect.

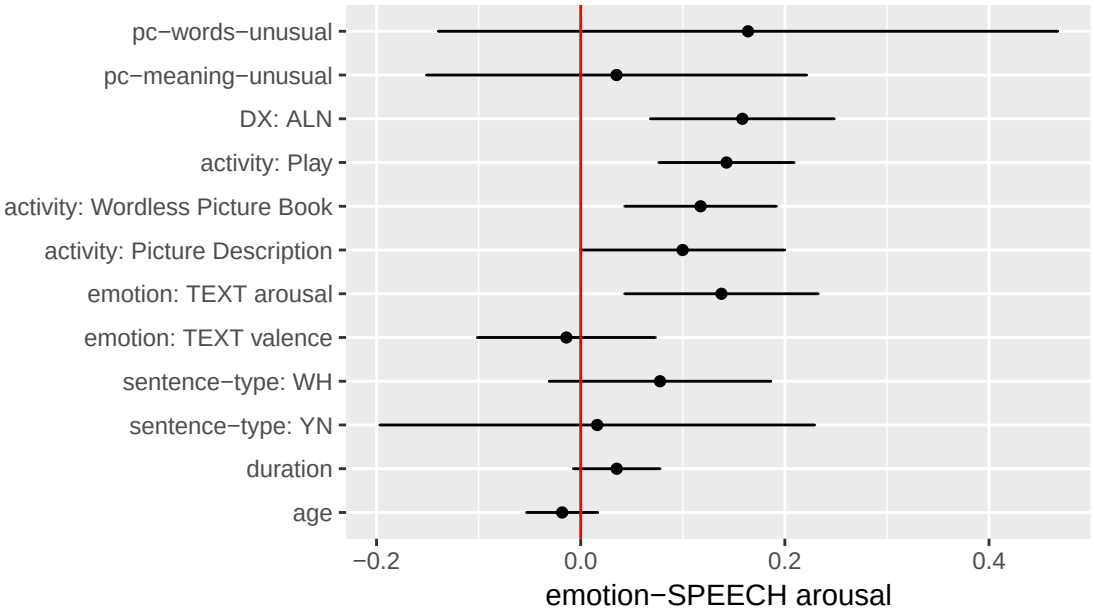


Figure 7.15: `MEP-S`: `SPEECH` arousal for the TD–ALN comparison: 95% confidence interval for the coefficients for the predictors in a linear model. The intercept corresponds to TD statements in conversations.

Figure 7.15 shows the confidence intervals for the coefficients for each factor. We can see that activity does have a significant effect, as well as sentence type, but the effect of DX remains very significant and larger for the `MEP-S` data set. The model explains only 9% of the variance (24% together with the random effects), so there are obviously many other factors that we do not know about. What is important here is that these content features are not responsible for the overall difference in arousal.

For the `MID-S` and `MID-D` data sets, the effect of the diagnostic group for predicting speech arousal approaches significance for ALN–TD, similarly to the results obtained using Monte Carlo tests; see Figures 7.16 and 7.17. It is not obvious from this data if this difference is due to some random factors, yet the agreement between the different data sets reinforces each finding.

We considered using transformed versions of certain predictors, but eventually decided not to use those for the emotional ratings. One may need to transform variables when they are used as predictors and highly collinear with each other, and when they are the outcomes examined and their distribution deviates from normality. An obvious disadvantage of doing these transformations is that the linear model becomes harder to interpret. This is aggravated if the parameters for the transformation are derived from the data itself, as is the case for example for PCA or Box–Cox

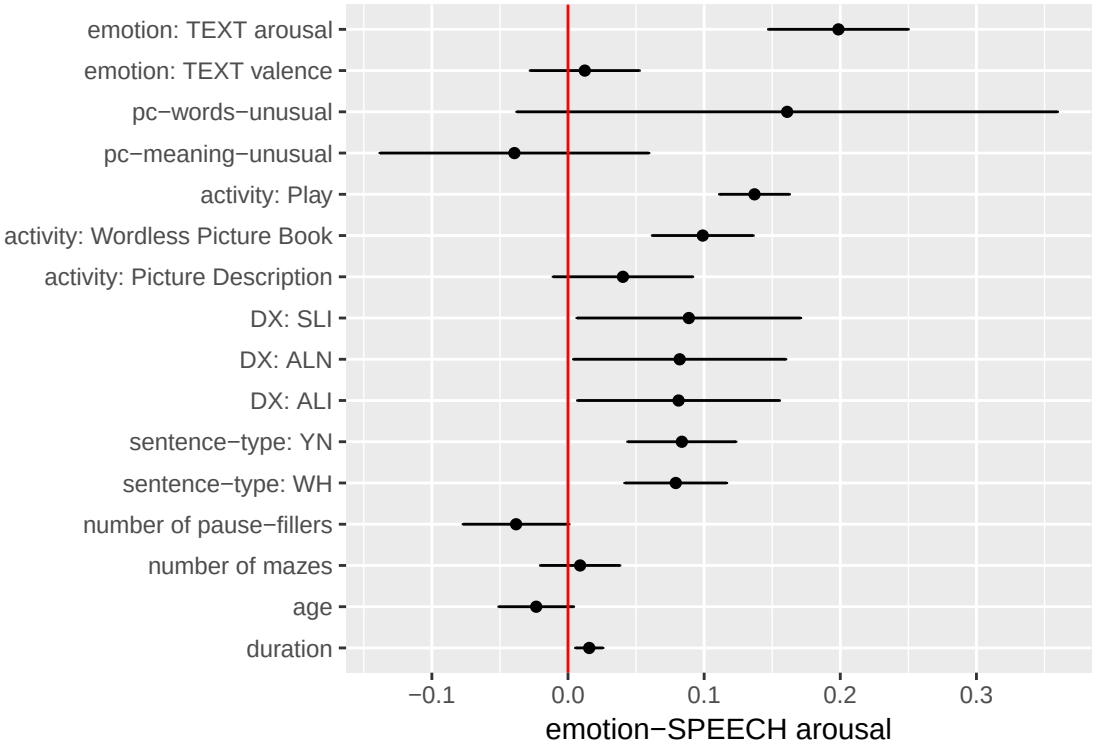


Figure 7.16: `MID-S: SPEECH` arousal for the TD–ALN comparison: 95% confidence interval for the coefficients for the predictors in a linear model. The intercept corresponds to TD statements in conversations.

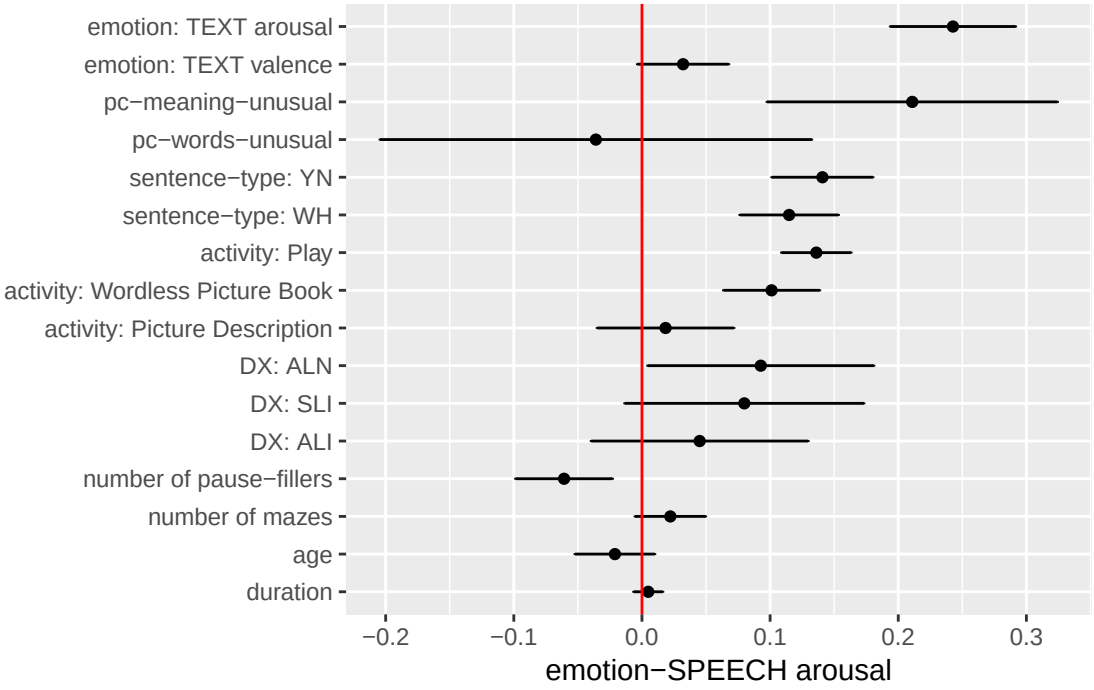


Figure 7.17: `MID-D`: `SPEECH` arousal for the TD–ALN comparison: 95% confidence interval for the coefficients for the predictors in a linear model. The intercept corresponds to TD statements in conversations.

transform, because when working with several data set, as is the situation here, the transformation parameters will inevitably vary from case to case, making the comparison of the analyses even harder. Working with the original variables avoids the above issues and thus is preferable when the issue that the transformations would treat is not grievous. Regarding the case of arousal and valence: When arousal played the role of the dependent variable, the optimal power for the Yeo–Johnson power transformation (see Section 3.4) varied widely from data set to data set (e.g. -0.3, $+0.5$, $+0.7$). The distribution skewness values did in fact become values close to zero using these, but a visual inspection of the original distributions did not indicate a large deviation from normality. Speech arousal and valence are also moderately correlated in all data sets, yet the models described above were able to determine in most cases that the `SPEECH` arousal rating is affected by the `TEXT` arousal rating, but not by `TEXT` valence (with the exception being the `MID-D` data set). This seems to indicate that the correlation between these variables is due to the distribution of emotions in the corpus and not a confusion of the meaning of these variables; see more about this in Section 7.3.3). Therefore we used these emotion ratings without transforming them.

"I don't know" answers

The `S` rating interface made it possible for raters to specify an "I don't know" rating for the emotional dimensions; the question is whether handling this type of ratings requires special attention. In our analyses of the arousal ratings, we simply ignored such ratings. We needed to exclude the possibility that their number distinguishes the groups.

It turned out that raters very seldom chose this answer, and it does not show a consistent difference between groups. In the `MEP-S` data set, it occurred in 0.2% of the arousal ratings, in the `MID-S` data set, in 0.5% of the cases. All groups had around the same percentage, except that the speech of SLI children were never rated this way in `MEP-S`; but they were in `MID-S`, so this is not a consistent difference either. Since the total number of these ratings is so small, it is not possible to find any consistent difference.

Confidence ratings for emotions

The equivalent of "I don't know" answers in the `D` interface was that the raters indicated their confidence in their answers, which might differ systematically by diagnostic group. We fit a mixed effect linear model to the `SPEECH` emotion confidence rating, using as predictors content features, the `TEXT` confidence rating, and the `SPEECH` emotion ratings, and then removed the non-significant ones. Transforming the confidence ratings to be approximately normally distributed and decorrelating the content unusuality rating or the arousal–valence pair did not help the model in any way, and as it would make it harder to interpret, we decided not to perform these transformations.

The marginal and conditional coefficients of determination for the minimal model were 28% and 31% respectively. See Figure 7.18 with the results. All predictors were approximately in the $[0, 1]$ range).

In summary, raters were more confident about their emotion ratings when the emotions were stronger and less confident when the prosody sounded atypical. The confidence emotional ratings for the textual content (provided by different raters) also had predictive value. No content features available to us, such as the activity or the sentence type, had a significant effect. Most importantly, the diagnostic group did not have a significant effect on the confidence of the emotion ratings. This does not necessarily mean that the raters were able to perceive the emotions of the speakers equally well when the speakers were autistic subjects, but being blind to the purpose of the study, at least they felt that they could when hearing the utterance in isolation.

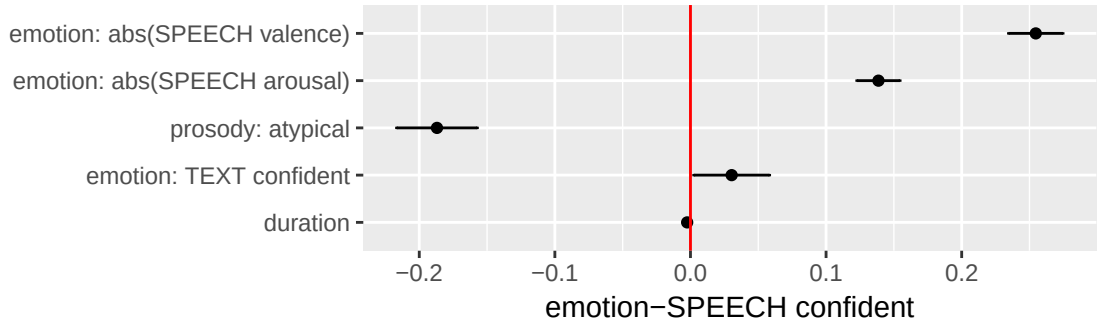


Figure 7.18: `MID-D: SPEECH` emotion confidence rating: 95% confidence interval for the coefficients for the predictors in a mixed effect linear model. The intercept corresponds to TD statements in conversations.

Arousal rating interpretations

Even though arousal seems to be a simple construct, there are some ways the raters might have misinterpreted it. Potential issues include associating it with positive feelings only when we asked where the emotion lay between the extremes of "calm" and "excited". It could also be misinterpreted as the intensity of the emotion, in which case it would not be independent of the level of valence but highly correlated with its absolute value.

We hypothesized that there are two main types of raters: one that interpreted arousal correctly, and one that misinterpreted it as the intensity of the emotion. The difference would show up for utterances with a negative valence, where for the former type of rater, arousal would be highly correlated with the absolute value of valence. Otherwise, the correlation of the two depends highly on the kind of emotions that appear in the corpus. In the case of children and the ADOS task, we believe that negative emotions with high arousal (such as angry, upset, or distressed) are unlikely to occur. So we expected that the arousal–valence correlation should be non-positive when the concept of arousal is interpreted correctly. We calculated the correlations of arousal with valence and its absolute value, and checked if raters formed clusters in the two-dimensional plane determined by these. We found no evidence of this.

Other evidence against the supposition that raters misinterpret arousal as the intensity of the emotion comes from its use in linear models. Consider the following examples, each of which counters this suspicion. First, the absolute value of arousal (i.e., its intensity) in the model for the emotion confidence ratings (see above in Section 7.3.3) is similarly important to the intensity of the valence rating, so the former cannot be equivalent to the latter. Second, raters agree much less about arousal when estimating it from text than from speech, whereas they are able to predict

valence from text well, and thus its intensity also (see Table 7.2).

One phenomenon that could be interpreted as supporting the conjecture that arousal, a concept orthogonal to valence, was misinterpreted as the intensity of the emotion is that it is moderately correlated with valence for `SPEECH`; see their correlations in Figure 7.19. Note though that this is not the only possible explanation. Another possibility is that this correlation is the result of the types of emotions displayed in the corpus utterances. Of course, one can specify the label for practically any coordinate in the valence–arousal plane, but it does not mean that these are equally likely, especially in a children's speech corpus. Indeed, most (over 35%) of the speech emotional ratings can be categorized as "happy", "excited", or similar (i.e., positive valence and high arousal), the next most frequent ones (23% each) are "content" or "gloomy" (low arousal and positive or negative valence), and the least frequent ones (9%) are "distressed", "angry", and similar (negative valence and high arousal); for an illustration of the emotional dimensions, see Figure A.1. The distribution of these ratings in the coordinate system in itself is enough to make the correlation high for `SPEECH`. For `TEXT`, arousal is harder to determine, and raters guessed for example that the child was "distressed" or "angry" about two times as frequently (18%) based on the text than based on speech, which in itself lowered the arousal–valence correlation for `TEXT`.

Another piece of evidence that arousal was indeed considered by raters to be orthogonal to valence is that linear models for speech arousal as the dependent variable were able to figure out that it is related to the text arousal rating but not to the text valence rating for all but one of the data sets, the exception being `MID-D`; see Figures 7.15, 7.16, and 7.17. Even for the `MID-D` rating set, the coefficient for text valence is much smaller. But it is indeed possible that the two ratings contaminated each other here, perhaps because raters chose the two ratings at the same time, with one click in a coordinate system. Having to indicate the answers with two clicks results in a temporal separation between the two decisions, which may have resulted in better results in this regard when using the `S` interface. On the one hand, the `D` interface gave raters a visual clue that the two dimensions are orthogonal.

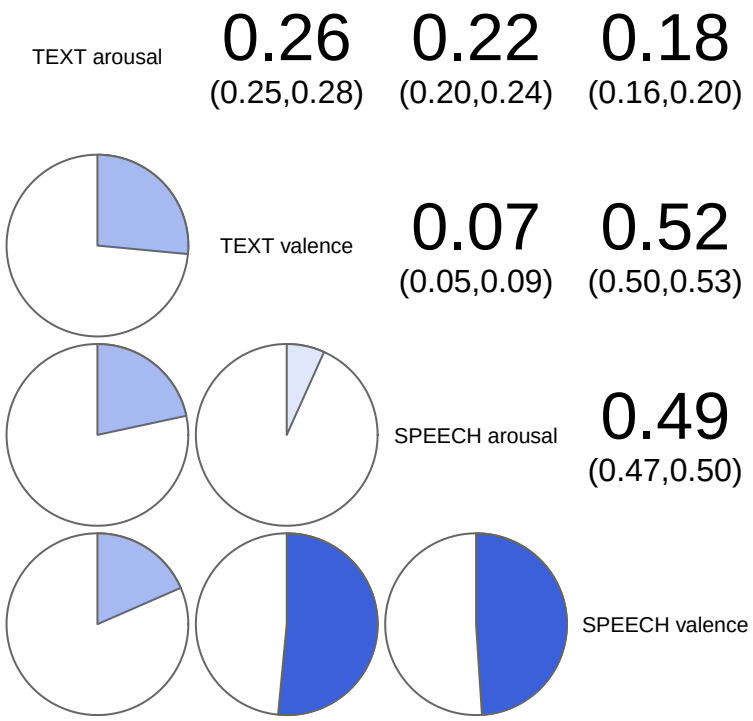


Figure 7.19: The correlation of the aggregated arousal and valence scores in all perceptual ratings data sets combined, with 95% confidence intervals

7.3.4 Prosody–content relationship

Significant group differences in ratings could be due to differences in content. Similarly, it is possible that after we control for content features, the effect of group membership on the ratings comes out as significant. To examine this possibility and to better understand what content features influence the raters' answers, we analyzed the perceptual ratings using mixed effect linear models (see Section 3.4.2).

We examined the speech prosody-related variables, namely prosodic atypicality, incongruency, arousal, and valence from the `SPEECH` task. The categorical predictors were the diagnostic group, the sentence type (Statement, WH-question, or YN-question), the activity (Conversation, Play, Picture Description, Wordless Picture Book; see Section 4.2.1). The intercept always corresponds to the statements in the conversations of the TD group. The numeric dependent variables were the utterance duration, number of mazes and of pause-fillers (these are only relevant for the `MID` stimulus set), and also the ratings for the `TEXT` task. We summarized the model outputs for all data sets separately for each variable in Tables 7.7, 7.8, 7.9, and 7.10. In the `MID` data set, all diagnostic group pairs are matched on relevant variables, as described in Section 4.2.4. In the `MEP-S` data set, only the ALN–TD and SLI–ALI groups are matched, and we chose utterance pairs for these matched groups separately. Therefore it would be inappropriate to compare TD with SLI, or TD with the merged HFA group (ALN + ALI) for `MEP-S`. Accordingly, we show the results for the ALN–TD comparison separately. We did not include the SLI–ALI comparison because there the diagnostic group is never a significant predictor.

The variance explained by the models is moderate to low (marginal R^2 from 0.09 to 0.34, conditional R^2 from 0.09 to 0.44). It is highest for valence, which is not surprising given that we use content features as predictors and we have seen earlier that the agreement for human ratings of text is highest for valence (see Table 7.2). Mostly, however, the amount of explained variance is much lower, so finding a predictor with a highly significant effect does not mean that it is responsible for a substantial amount of the variation in the ratings.

A common element in the models is that utterance duration is positively associated with each rating. The reason must be that it is easier to detect the presence of any speech phenomenon from longer speech segments. When the utterance is short, raters cannot confidently say that the utterance either sounds atypical or very emotional.

One caveat is that some groups of variables were highly correlated and our way of addressing that in the mixed effect linear models affects the interpretation of the models. We have discussed the case of arousal and valence earlier: These do not cause a problem during the analyses so we

did not transform them. It was not clear however that the unusuality ratings were not causing artifacts in linear models, as they sometimes had opposite and high coefficients. We addressed this by replacing the highly correlated variables with their principal component scores, denoting each one with the "pc" prefix and naming it after the rating that it is most highly correlated with for easier interpretability: `pc-meaning-unusual` and `pc-words-unusual`. Note that we derived them separately for each rating data set, so they may not mean exactly the same thing for different data sets. Below we summarize what we can see from the model coefficients.

The speech arousal–content relationship

As we can see in Table 7.7, there are several reliable predictors for speech arousal across the data sets. Most importantly, the ALN, SLI, and ALI groups always have a higher arousal than the TD group, which the intercept belongs to, with ALI only approaching significance for `MID-D`.

One of the best predictors is the arousal rating determined from text. The second most important one for the `MID` stimulus set is an unusual content rating; note that the principal components of the unusuality ratings are used, which are correlated with both "meaning unusual" or the "words unusual" ratings. The activity also has a significant effect, with higher arousal especially in the play-situation, and all others as well compared to conversations. Questions are also generally associated with higher perceived arousal. The presence of mazes is positively correlated with arousal, the presence of pause-fillers is negatively correlated with it. Presumably when the children are more excited, they generally try to express themselves without careful planning, thus making more errors.

The speech valence–content relationship

The best predictor for speech valence is the valence rating determined based on the text. The only other consistent difference is that each activity is associated with higher valence (i.e., more happiness) than Conversation (which in the ADOS is sometimes about difficult topics designed to uncover issues associated with autism). A diagnosis of either HFA or SLI is not significantly associated with a difference in this rating; see Table 7.8.

The atypical prosody–content relationship

For the `MID` stimulus set, but not for the `MEP` data set, both HFA and SLI is associated with higher prosodic atypicality scores; see Table 7.9. There are no other consistent differences across the data sets.

Prosody–content incongruity

For the MID-D data set, the groups with an HFA or SLI diagnosis are associated with somewhat higher perceived incongruency between the prosody and the content of their utterances, whereas for the smaller (both in number of utterances and in number of ratings per utterance) MID-S data set only the TD–ALN difference reaches significance; see Table 7.10. Similarly to prosodic atypicality, there is no effect of diagnostic group for the MEP stimulus set.

Table 7.7: SPEECH arousal: coefficient values from a MELM model. The arousal and valence predictors are for TEXT. For each predictor, its coefficient is given with its significance level for differentiating the groups is marked after its name (*** 0.001 ** 0.01 * 0.05 + 0.10).

MEP-S TD–ALN	MID-S	MID-D
marginal R2: .09	marginal R2: .09	marginal R2: .12
cond. R2: .24	cond. R2: .25	cond. R2: .29
ALN**: .16 Play***: .14	arousal***: .2 valence+: .01	arousal***: .24 valence*: .03
Wordless Pic-	pc-words-unusual*: .16	pc-meaning-unusual***: .21
ture Book**: .12	pc-meaning-unusual+: -.04	YN***: .14 WH***: .11
Picture Description*: .1	Play***: .14 Wordless Pic-	Play***: .14 Wordless Pic-
arousal**: .14 WH+: .08	ture Book***: .1	ture Book***: .1
duration+: .04	Picture Description*: .04	Picture Description+: .02
	SLI**: .09 ALN**: .08	ALN*: .09 SLI*: .08
	ALI**: .08 YN***: .08	ALI+: .05
	WH***: .08	Pause-Fillers***: -.06
	Pause-Fillers*: -.04	Mazes*: .02 age*: -.02
	Mazes+: .01 age*: -.02	duration+: 0
	duration***: .02	

Table 7.8: SPEECH valence: coefficient values from a MELM model. The arousal and valence predictors are for TEXT. For each predictor, its coefficient is given with its significance level for differentiating the groups is marked after its name (*** 0.001 ** 0.01 * 0.05 + 0.10).

MEP-S TD–ALN	MID-S	MID-D
marginal R2: .32	marginal R2: .21	marginal R2: .34
cond. R2: .41	cond. R2: .28	cond. R2: .44
valence***: .44	valence***: .39 YN***: .07	valence***: .41
pc-meaning-unusual*: .13	Wordless Pic-	arousal***: .05
pc-words-unusual+: .13	ture Book***: .07	pc-meaning-unusual***: .28
WH**: -.09	Picture Description**: .07	YN***: .08 WH***: .05
Wordless Picture Book*: .06	Play***: .06	Play***: .08 Wordless Pic-
Picture Description+: .05	Pause-Fillers*: -.03	ture Book***: .07
Play*: .04 ALN+: .03	Mazes*: -.02	Picture Description*: .04
	duration***: .03	duration***: .01

Table 7.9: SPEECH atypical: coefficient values from a MELM model. The arousal and valence predictors are for TEXT. For each predictor, its coefficient is given with its significance level for differentiating the groups is marked after its name (*** 0.001 ** 0.01 * 0.05 + 0.10).

MEP-S TD–ALN	MID-S	MID-D
marginal R2: .15	marginal R2: .09	marginal R2: .16
cond. R2: .19	cond. R2: .12	cond. R2: .23
valence***: -.11	duration***: .03 ALI**: .03	pc-words-unusual**: -.11
duration***: .1 Play*: .03	ALN**: .03 SLI*: .02	pc-meaning-unusual+: .03
	WH*: -.02 Play**: .02	duration***: .05 SLI**: .04
		ALI**: .03 ALN*: .02
		Mazes***: -.03
		Pause-Fillers*: -.01
		WH**: -.02 YN*: -.02
		valence*: -.01

Table 7.10: SPEECH incongruous: coefficient values from a MELM model. The arousal and valence predictors are for TEXT. For each predictor, its coefficient is given with its significance level for differentiating the groups is marked after its name (*** 0.001 ** 0.01 * 0.05 + 0.10).

MEP-S TD–ALN	MID-S	MID-D
marginal R2: .15	marginal R2: .06	marginal R2: .09
cond. R2: .19	cond. R2: .09	cond. R2: .18
valence***: -.17	duration***: .02 WH**: -.02	pc-words-unusual***: -.1
arousal***: .07 WH**: .07	ALN*: .02 SLI+: .01	pc-meaning-unusual***: .09
duration***: .04		duration***: .02
Wordless Picture Book+: -.03		WH***: -.02 YN*: -.01
		SLI*: .02 ALI*: .01
		ALN*: .01 Mazes***: -.01
		arousal*: .01

7.3.5 Atypical aspects of speech prosody

The SPEECH ASPECTS ratings show group differences in accordance with what one would expect based on our a priori knowledge about autism, but the differences are significant only for the MEP data set. For each utterance, we calculated the mean of all the atypical speech prosody aspect ratings, namely for loudness, pausing, pitch, speed, stress, and voice quality. We derived the per-subject means for these, and the per-DX averages of the per-subject means. We can see the result in Figures 7.20 and 7.21 for the MEP-S data set, and in Figures 7.22 and 7.23 for the MID-D data set.

Based on the figures, it seems that the raters on average gave a smaller number of atypical speech aspect ratings to the TD group. We compared the average of the atypicality ratings per utterance for the SLI–ALI and ALN–TD comparisons, but the difference was not significant in either case. Note that the number of utterances rated was less than 500 in each data set.

We calculated the first six principal components (PC) of all the aspect ratings, and compared them across the groups. Only for the MEP stimulus set, we found a significant difference ($p < 0.01$) using Monte Carlo testing for these PC components together. Testing the PCs separately, only the first one (PC1) came out as different (uncorrected $p < 0.02$), and the second one (PC2) approached significance; see Figure 7.24. PC1 was associated most strongly with too soft speech, too low pitch, and negatively associated with too effortful speech and too varied or too high pitch; this type of

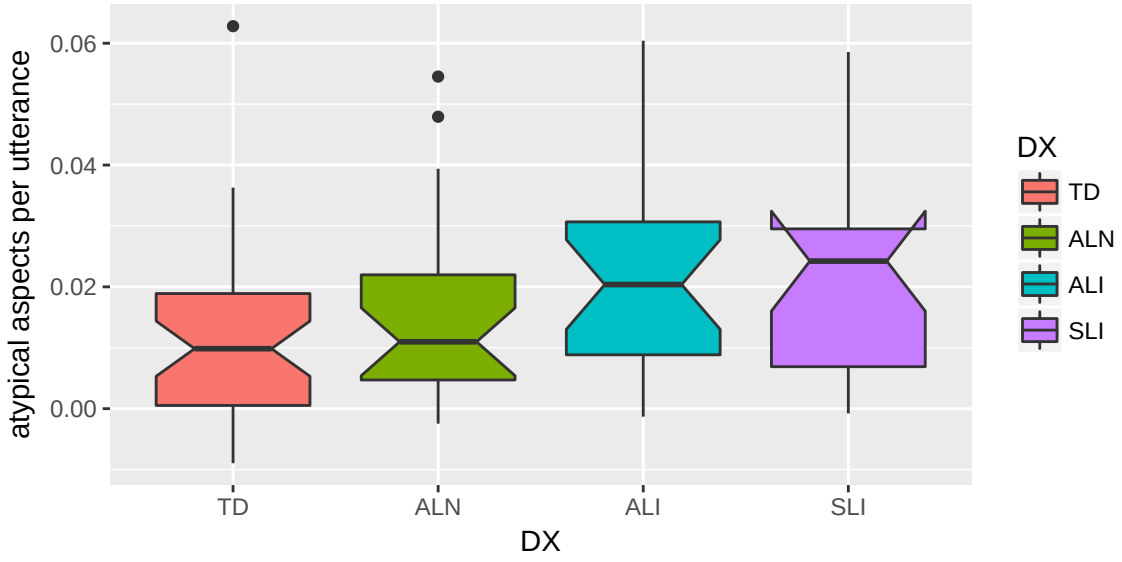


Figure 7.20: `MEP-S`: The per-subject averages for the mean number of atypical speech aspects per utterance

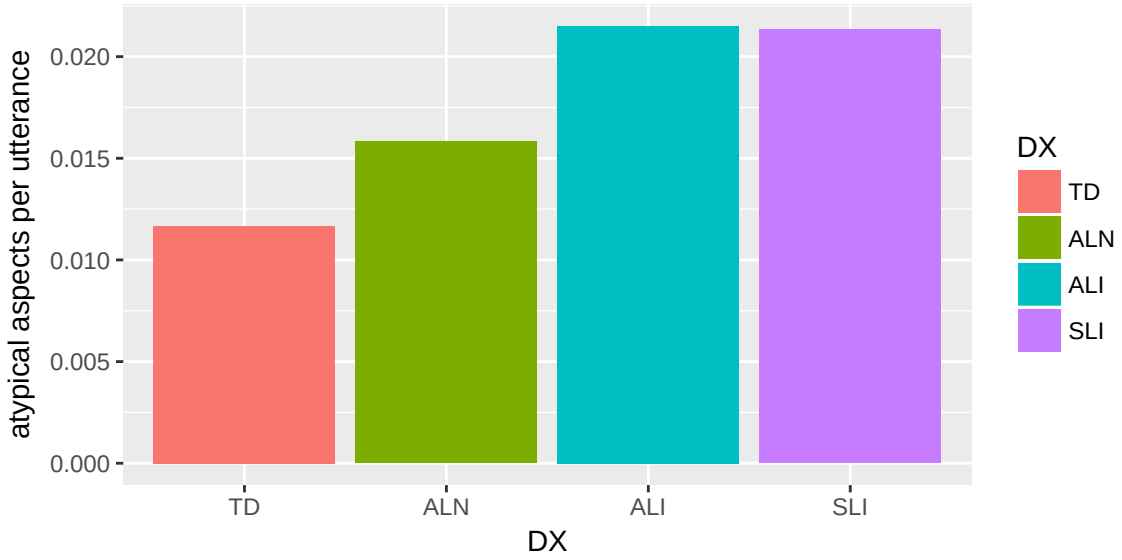


Figure 7.21: `MEP-S`: The per-DX averages of the per-subject means for the number of atypical speech aspects per utterance

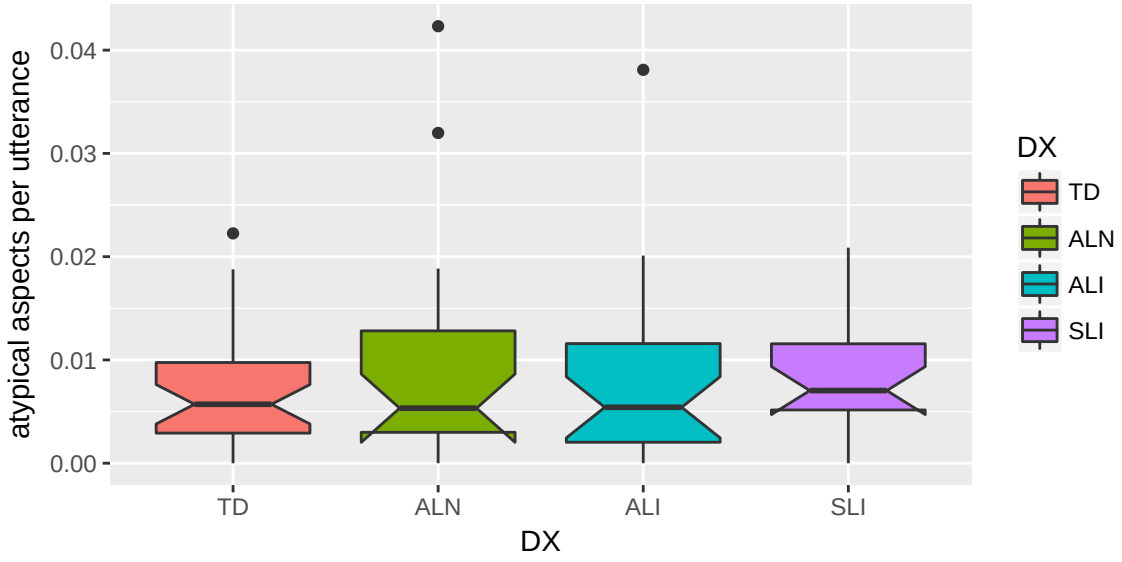


Figure 7.22: `MID-D`: The per-subject averages for the mean number of atypical speech aspects per utterance

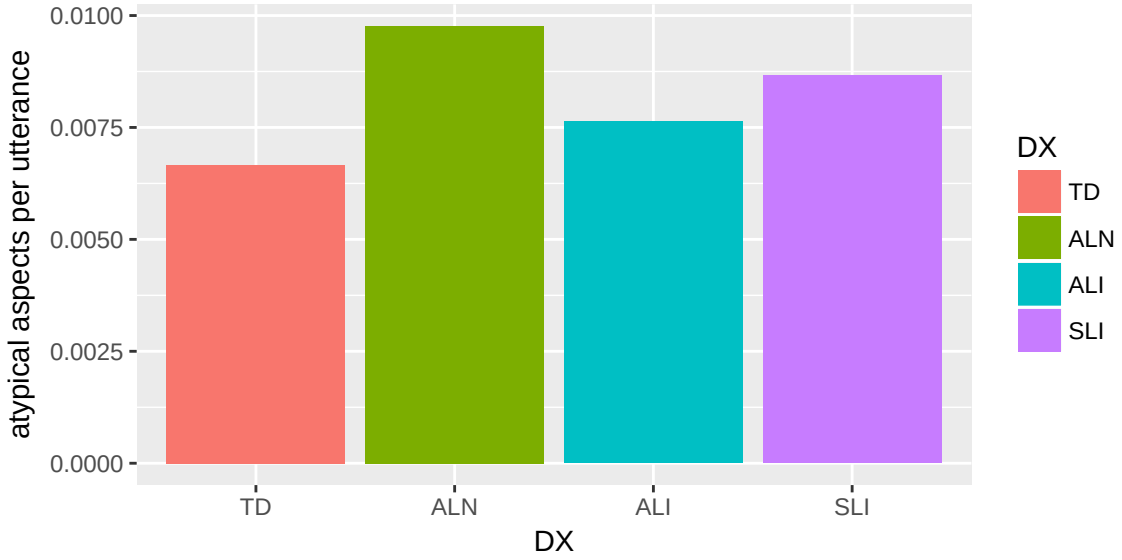


Figure 7.23: `MID-D`: The per-DX averages of the per-subject means for the number of atypical speech aspects per utterance

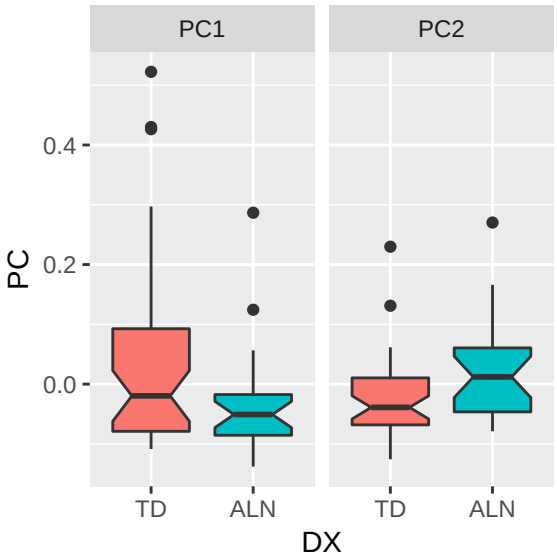


Figure 7.24: `MEP-S` TD–ALN comparison: Per-subject means of the first two principal components for the `SPEECH ASPECTS` utterance ratings

atypicality occurs more for TD children. PC2 was associated most strongly with wrong pausing strategy, misplaced stress, too high or too varied pitch, too varied or too slow speed, and too effortful or too varied speech loudness; this occurred more for children with ALN. These principal components do not separate the two groups reliably, as we can see in Figure 7.25.

Applying Fisher's linear discriminant analysis to the data, we got very similar results. The coefficients with the highest positive difference for ALN in decreasing order of the magnitude were: pitch too high, loudness too effortful, pitch too varied, stress bad, loudness too varied, voice quality very tense, pausing wrong, and speed too varied. The coefficients with the highest negative difference for TD in decreasing order of the magnitude were: loudness too soft, pitch too low, and pitch too flat. Again, this projection of the features was able to classify the utterances correctly in only 55% of the cases, which is not much better than chance.

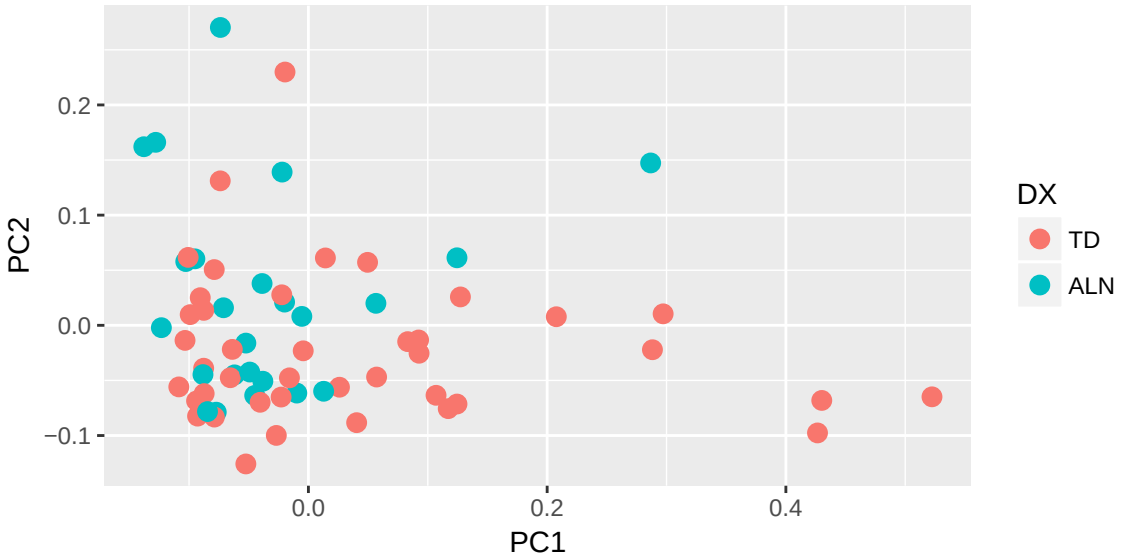


Figure 7.25: `MEP-S` TD–ALN comparison: Per-subject means of the first two principal components for the `SPEECH ASPECTS` utterance ratings

7.4 Discussion of the Differences in Subjective Ratings

In light of the stimulus selection process, the rationale behind the task interfaces, the rating collection, and the analyses of the collected perceptual ratings, a pattern of findings seems to emerge. First, the way the stimuli were chosen had a significant effect on the speech prosody findings. Second, we found some of the expected differences between autistic and the typically developing children, but the thousands of utterances and tens of thousands of ratings often resulted in only barely significant overall differences between statistical properties of the groups, and very small effect sizes. This indicates that, at least for these high-functioning children, the atypicality is present only in a subset of their utterances. Third, we did not find consistent prosodic differences between children with autism and those with specific language impairment: children with SLI displayed similar kinds of speech atypicalities to children with HFA.

It is important to point out that neither the stimulus set matched on expected prosody (`MEP`) nor the maximally individually diverse (`MID`) stimulus set was constructed to be representative of the overall distribution of the properties of each subject's speech. So when we do not find a difference to be significant between two groups, it does not mean that uniformly sampling the conversations of the same children would not result in a difference. The data set we collected does however have the capability to be used for this kind of analysis as well. For example, it can be oversampled based on a match between content features and acoustic-prosodic properties of the

utterances in the speech corpus and those in the ratings data set; any differences derived from this oversampled rating set likely generalize to the speech of each subject. As it is, we need to keep in mind the properties of the stimulus sets when interpreting the findings associated with them.

The MEP stimulus set was matched on content features as described earlier, including the part-of-speech sequence, the particular words whenever possible, the emotions expressed by the words, and the unusuality of the content, and it included no utterances with mazes or grammatical errors. Without formal testing, based on the fact that each of these utterances occurred in the speech of several children, or at least very similar utterances were present, we can conjecture that the chosen utterances were among the most frequent, most typical ones even for the children who otherwise produced much erroneous or unusual content.

The purpose of the MID stimulus set was to cover the whole repertoire of utterances with diverse content and prosodic features separately for each subject. Therefore it is bound to sample the more atypical content when the subject has any. It is also more likely to contain new, creative content. This is supported by the fact that the difference in the "meaning unusual" rating approached significance for the MID, but did not differ at all for the MEP stimulus set. The picture is not totally clear though, as the "words unusual" ratings were high both for MEP-S and MID-D, and less so for MID-S. Nevertheless for the discussion below, we conjecture that the MEP stimulus set contains more typical verbal content than the MID stimulus set does. Let us review the findings in light of these qualities of the stimulus sets.

Certain ratings showed different patterns between the our stimulus sets: For the MEP set, we found very pronounced differences in the SPEECH arousal ratings between TD and the atypical groups, but no differences in the prosodic atypicality ratings. For the MID set, the situation was different in that the arousal difference was less pronounced, whereas the atypicality ratings differed significantly for the MID-S data set, and for both MID data sets after controlling for content features.

This phenomenon seems to indicate that content has a very pronounced effect on these ratings. There are three main interpretations for this, all of which may be true at the same time: One is that children with HFA can produce typical content with more typical prosody. The other is that listeners tend to overlook somewhat atypical prosody or to interpret it as an expression of emotions when the utterance content is typical. The third is that children with HFA experience more arousal in the situation staged by administering the ADOS. Let us look at each possible explanation in detail.

1. It is possible that children with HFA and SLI uttered the sentences with the typical content — such as those in the MEP stimulus set — in an equally typical way to their TD peers. If a

sentence is fairly frequent in everyday speech, then children hear many instances of them with typical prosody. If they are capable of imitating intonation patterns, then it is easier for them to produce such utterances with appropriate prosody just using their imitation abilities. On the other hand, it may be harder for children with neurodevelopmental disorders to produce appropriate prosody for a new sequence of words that they come up with on the fly. The classical case of the autistic child as described by Kanner (1943) does not use speech for communicative intent until being much older, and when he starts doing that, it is usually using "canned utterances" first. Even though the diagnostic category of autism has been extended to include children for whom this is far from being the case, it is reasonable to assume that it is harder for some children with ASD to synthesize new content, especially putting all aspects in place on the fly, including prosody.

2. Another possibility is that the typical content prompted the raters to overlook any minor atypicalities in prosody. Based on our analyses with the mixed effect linear models, the "meaning unusual" and "words unusual" ratings did not have a consistent effect on the prosodic atypicality ratings, but they are correlated with the arousal ratings on the `MID` data set. We did not find this correlation for the `MEP` data set. Indeed we cannot because the "meaning unusual" rating is much rarer there, but the difference in speech arousal ratings was more pronounced. Therefore it seems likely that raters attributed prosodic atypicalities for typical content to the child being more excited or annoyed. We must remember here that they heard isolated utterances from the children, so they may perceive some of these utterances as more atypical if presented in context.

3. The emotions identified from speech may not be the actual emotions experienced by the speaker, but there is evidence that children with autism indeed experience more arousal in social contexts. O'Haire et al. (2015) measured physiological signals associated with arousal on children with autism and typical development in varying contexts (reading silently, reading aloud, free play, and playing with animals). They found that children with ASD experienced arousal in a different pattern across contexts than their TD peers, and in social contexts they experienced higher arousal. Based on this, it is highly likely that our subjects with HFA were not only perceived as being more excited, but indeed experienced more arousal than the children in the TD control group. Interestingly, the SLI group showed approximately the same level of speech arousal as the HFA group.

Atypical prosody — considered so by the majority of raters — occurred in typical speech as well. This is in accordance with the observation of Peppé (2009), among others, who stated that

"it is nearly always possible to find prosodic rules transgressed in perfectly normal usage, and conversely ... when characterizations of atypical prosody are made, it is nearly always possible to find examples of them in typical speech."

The raters were equally confident in their emotion ratings irrespective of the diagnostic status of the child being rated. Our analyses using mixed effect models showed that higher prosodic atypicality decreases the rater's confidence in his or her emotional judgement, while a higher level of emotional intensity increases it. Since children with HFA sounded more atypical on average, if this was the only factor at play then there would be an air of uncertainty about what kind of emotion they are experiencing. It seems that this is not the case because they generally display a higher level of emotional intensity, particularly increased arousal. Their perceived valence ratings are not significantly different from the TD group however, so raters may be less sure about their valence ratings. To see if this is so, we compared the standard error of the valence and arousal ratings across the groups. As expected, the standard error for valence was higher for each group than that for TD. After controlling for relevant content features similarly to our earlier analyses, we got an uncorrected $p < 0.05$ for the TD–ALN difference only, but not for the other group pairs. The difference in the standard error of the arousal ratings was the opposite, but the differences between the group pairs were not significant.

Congruency ratings did not differ between the groups significantly. The mixed effect model coefficients associated with the diagnostic categories did point in the expected direction, with both HFA and SLI being associated with less congruous relationship between prosody and content. These ratings had the lowest agreement between the raters, probably partly because there is some inherent ambiguity in how it relates to the prosodic atypicality and unusual meaning ratings. When prosody and content are discrepant, some raters may have marked it as having atypical prosody, others as being incongruous, depending on whether they gave precedence to the content or the prosody when interpretating the utterance. In similar experiments, it may be necessary to explain various cases in detail or to leave this question out.

Utterance duration is positively correlated with all ratings. It is presumably because it is easier to detect any speech phenomenon from longer amounts of speech, and thus raters indicate the presence of these only when there is enough evidence.

Based on the outcomes of our analyses, we can synthesize some lessons learned for future studies. We collected a uniform number of five ratings for each utterance for `MEP-S`, and at least three for each utterance for `MID-S` and more for utterances that had a high standard error in the ratings. We had consistently higher rater agreement for all ratings using the latter approach, even though the average number of ratings per utterance was well below five. So the ratings collected

this way were not only better, but also cheaper. It also seems possible that asking for the valence and arousal ratings in two separate steps helps the raters to treat them as orthogonal concepts.

Limitations of the study include that the size of the stimulus set with perceptual ratings is moderate, and neither stimulus set is representative of the speech of the individual subjects. Negative findings do not mean that the diagnostic groups under study do not differ in the particular measure. Similarly, a positive finding may not stand up when examined on a representative speech sample. The spontaneous speech corpus was collected using a particular methodology, the ADOS, and may not generalize to corpora collected using other methods. Our findings characterize each group as a whole and may not be true for every individual in their respective groups.

In preliminary analyses with the three score aggregation methods described in Section 7.2.9 — namely the simple averages, the z-normalized scores, and the correlation-based approach for estimating rater competencies that we introduced — we compared the results obtained using each one in mixed effect models. We got higher explained variance and more significant differences when using the correlation-based aggregation method than when using the other two. Note that each rating variable was aggregated independently of the other variables, so it cannot be an artifact of the methodology. One issue that must be dealt with is that the rater competence can become negative for some raters. It is theoretically possible that some raters misunderstand the direction of the scale and provide scores that are perfectly usable after they are reversed, and this correlation-based approach automatically deals with that by reversing such scores. Yet it may still result in unexpected score range changes, so it is worth considering excluding raters that are assigned negative weights. Another issue is that aggregating scores by weighing them using rater competences can make the scores grow outside the original ranges. For example, for a variable in the $[-1, 1]$ range, the aggregated score may be in the $[-0.8, 1.2]$ range. In such cases, the aggregate score may need to be transformed back to the original range.

7.5 Summary for the Perceptual Rating Collection and Analysis

We worked with a well-characterized spontaneous speech corpus of children with autism, and others with language impairment or typical development as control groups. We designed and executed a study that involved the collection of prosodic and emotional perceptual ratings as well as subjective content ratings for isolated utterances, and analyzed the resulting data using statistical methods.

We worked with two carefully chosen stimulus sets: First, utterance pairs were chosen to be matched on "expected prosody". Second, utterances maximally diverse with respect to both

content and prosody for each individual. The introduced correlation-based rater competence estimation method can deal with a varying number of answers per rater, seems to reflect external information and to result in more reliable score estimates than some alternative approaches. The resulting corpus of subjective ratings can be used as a basis for further research and for predicting perceptual ratings automatically. Our analyses concentrated on the prosody ratings while controlling for the content features, which may be unique among such studies.

We have found group differences consistent with earlier findings, including that HFA was associated with higher prosodic atypicality and higher perceived speech arousal. We discussed our findings and their relationship to the content features in detail, here we briefly review them. The rater agreements ranged from moderate to high for emotional ratings, and from moderate to low for prosodic atypicality and unusual content ratings using the presented methodology for collecting ratings. We consistently found greater arousal in HFA and SLI than in TD after controlling for content features, likely due both to them actually experiencing more emotional arousal and also to the listeners interpreting atypical prosody as a sign of emotionality, especially when the utterance content is typical. Listeners identified utterances with atypical prosody in typical development as well, but the kind of atypicality may be different: Utterances of TD children seem to be considered atypical most often because they speak too softly and monotonously. This cannot be due to varying audio volume, first, because we normalized the recording volumes (see Section 7.2.7), second, because we asked about the effort exercised by the speaker and human listeners can generally detect vocal effort quite well from speech independenly of the actual volume (Brandt et al., 1969). Utterances of children with HFA may be perceived as atypical most often due to wrong pausing, misplaced stress, and too varied prosody as well as too high pitch, too slow speed, and too effortful enunciation. Importantly, the language-impaired group did not differ on any measures significantly from a matched group of children who had both language impairment and autism.

Chapter 8

Perceptual Ratings of Prosody: Prediction from Acoustic Features

In the previous chapter, we described how we collected and analyzed a perceptual rating dataset for children's speech utterances. Now we go one step further by setting up machinery for predicting these ratings and ratings from another corpus available to us from acoustic-prosodic features. We first motivate this work, then go on to describe our results.

8.1 Motivation for Predicting Subjective Ratings

Earlier research (see e.g. Peppé, 2006; Nadig and Shaw, 2012; Filipe, 2014) as well as our work (see Chapter 7) has shown that various perceptual ratings of prosody and emotions differentiate autism from typical development to some extent. While this is a potentially useful finding for screening purposes, it often requires the aggregation of the opinions of multiple judges to get reliable estimates of speech characteristics; even trained clinicians lag behind in precision when they need to score their subjects real-time during an examination (see e.g. van Santen et al., 2009). Aggregated scores from multiple raters for recordings of the examinations can give more reliable scores, but this procedure is obviously time-consuming and costly, whereas automated methods have the promise of providing scores that are sufficiently precise much more conveniently (van Santen et al., 2009). The questions remain what approaches can predict perceptual ratings of prosody and emotions and whether currently those have sufficient accuracy to provide diagnostically relevant information.

We concentrate on providing automated methods for evaluating two measures that have been confirmed to be diagnostically relevant using perceptual ratings: First, Jan van Santen hypothesized based on what we know about functional and structural connectivity issues in autism (see e.g. Barnea-Goraly et al., 2004) that the correlation between the emotions in different modalities — including speech, gesture, and facial expression — is lower in autism than in typical development. He worked with valence and arousal ratings, which are two of the most important aspects in the dimensional model of emotion (see e.g. Scherer, 2005). Based on the perceptual ratings, he was able to confirm the hypothesis for valence (van Santen, 2014): Across multiple modality-pairs, including the speech–text pair, the average correlations between the per-subject valence scores was significantly higher in TD than in HFA. Second, we found (see Section 7.3.2) that utterances of children with HFA carry significantly higher arousal that those of children with TD when the utterances are matched on content features. This arousal difference approached significance for another stimulus set that was maximally diverse for each individual as well. For the latter stimulus set, the difference in atypicality ratings approached or reached significance as well.

In this chapter, we describe our work on predicting perceptual ratings of emotions and prosodic atypicality from acoustic features of speech prosody, with the aim of producing diagnostically

relevant automated features. Prosodic features, if they are internally valid and suitable for the task at hand, have the advantage compared to spectral features that they do not carry information about the speech coding and the channel characteristics or the vocal characteristics of the speaker. These are factors that complicate the task of speech recognition for example, and require such systems to do speaker and channel normalization. We worked with two different kinds of children's speech corpora, one containing acted emotions, and the other spontaneous speech to evaluate our approach. We base this discussion partly on a paper we published with co-authors Meysam Asgari, Izhak Shafran, and Jan van Santen (Asgari et al., 2014).

8.2 Methodology for Predicting Perceptual Ratings

8.2.1 Predicting emotions for the CSLU Cross-Modal Corpus

We worked with a corpus of children's recordings who acted out brief stories. The corpus contained 835 video recordings of 28 children. From the videos, five modality-specific versions were prepared for each sentence: speech, delexicalized speech, the transcript, face only, gesture only. Of these, we worked with the speech and the delexicalized speech recordings. The utterances were rated for emotional valence and arousal and an aggregate gold-standard score was derived. The agreement for arousal for the speech and delexicalized speech ratings was 0.88 and 0.85, respectively. For valence, the agreement was 0.91 and 0.75. Our goal was to predict these subjective ratings from acoustic-prosodic features. For more corpus characteristics, see Section 4.1.

Features

We extracted acoustic-prosodic features from the original and the delexicalized speech recordings. We automatically removed silence from the start and end of utterances, based on the absence of voicing or very low RMS values ($< 40\%$ percentile of the RMS of all unvoiced parts), keeping parts with large intensity ($> 95\%$ percentile of silence parts) as speech. We determined the F0 and intensity curves and transformed them into the log-domain for feature extraction as described in Section 3.1. We extracted two feature sets from these: statistical features, which only reflect properties of the distribution of the F0 and intensity samples disregarding their locations in the utterances, and functional features derived using fPCA (see Section 3.3), which make use of the curve shapes. The feature set comprised a total of 54 prosodic features.

Statistical features We calculated a total of 34 per-utterance robust and non-robust statistical moments for F0 and intensity. This feature set is based on our earlier work on characterizing

autistic speech (Kiss et al., 2012), also used in Chapter 6 of this dissertation. For both the F0 and intensity curves, we calculated 16 features, namely the variance and the features listed in Table 6.1. To help the model weigh the importance of higher-order moments, which require more input frames to be estimated reliably than lower order moments, we added the number of voiced frames and the total number of speech frames as features. We examined the effect of excluding location statistics of the intensity features (such as minimum, mean, median, etc.), assuming that these may reflect the distance of the microphone from the speaker rather than speaker characteristics. This step did not change the performance significantly.

FDA features We used Functional Data Analysis (FDA, see Section 3.3) utilizing the `fda` R package (Ramsay et al., 2014) to characterize the shape of the F0 and intensity curves. We calculated the first 10 fPCA coefficients for both F0 and RMS, resulting in 20 features per utterance. An advantage of this approach is that it takes all curves into account when determining the feature vector for a particular utterance. The coefficients with lower indices belong to the component curves that explain a higher amount of variance among all curves. (For example, fPCA coefficient 1 explains at least as much variance as fPCA coefficient 2.)

Training and evaluating the regression model

We used Support Vector Regression (SVR) with an RBF kernel for predicting the ratings. For evaluation, we used a five-fold cross-validation scheme: We divided the training set into five subsets. We set all model parameters based on four of the five subsets and assessed the model performance on the fifth ones. For arousal, we determined the cost and γ meta-parameters using a grid-search in a cross-validation scheme on the training set. For valence, we always used the default values in the e1071 R package (cost = 1 and $\gamma = 1$ divided by the number of features) because grid-search did not improve performance.

8.2.2 Predicting perceptual ratings for the CSLU ERPA ADOS Corpus

We trained and evaluated models for predicting four prosody-related ratings for the utterances, namely arousal, valence, prosodic atypicality, and incongruency. We used the `MID-D` data set described in Section 7.2.4 and derived the gold standard scores by calculating the weighted mean of the z-scores of ratings from multiple raters (see Section 7.2.9).

Features

We extracted acoustic-prosodic features from the speech recordings of the CSLU ERPA ADOS Corpus. The utterance boundaries were known from human transcriptions. We determined the F0 and intensity curves and transformed them into the log-domain for feature extraction as described in Section 3.1. The feature set we worked with was special in that it contained not just per-utterance features, but also per-speaker features from the whole corpus. It is a reasonable assumption that this information is available if we have a substantial amount of speech recorded for the subjects and we know the identity of the speaker, which is generally the case. The feature set comprised 30 per-utterance and 30 per-speaker variables, namely statistics of the F0 and intensity curves; see Table 6.1.

Training and evaluating the regression model

We used Support Vector Regression (SVR) with an RBF kernel for predicting the ratings. For evaluation, we used a leave-one-speaker-out cross-validation scheme: We trained the model on the data of all speakers but one, then predicted the ratings for the speaker left out. We determined the cost and γ meta-parameters using grid-search in a 10-fold cross-validation scheme on the training set using the `tune` function in the `e1071` R package (Meyer et al., 2017).

8.3 Prediction Performance Results

8.3.1 Predicting emotions for the CSLU Cross-Modal Corpus

Table 8.1: Correlation between the gold standard and the predicted emotion ratings. We predicted the ratings using an SVR model and various acoustic-prosodic feature sets.

	Number of features	Arousal	Valence
Delexicalized Speech			
FDA	20	0.50	0.31
Moments	34	0.78	0.35
Moments+FDA	54	0.77	0.39
Human ratings		0.85	0.75
Speech			
FDA	20	0.59	0.22
Moments	34	0.83	0.28
Moments+FDA	54	0.83	0.34
Human ratings		0.88	0.91

We trained and evaluated the regression model in a five-fold cross-validation scheme as described earlier, and report the average of these performance estimates in Table 8.1. Our goal was to recognize these emotions from speech with enough precision to be able to estimate the cross-modal correlations between speech, delexicalized speech, and text. The question is whether this precision is enough. This work was done as a team, and we need to rely on the result of our colleagues as well to evaluate this.

In the course of this work, Meysam Asgari used another feature set for the same data set, which contained prosodic features based on a harmonic model of speech as well as spectral features (Asgari et al., 2014). The results using this feature set were around the same as ours for arousal, and better for valence (0.47 for speech and 0.42 for delexicalized speech). The combination of the two feature sets for the speech task improved the performance further to 0.86 for arousal and 0.53 for valence. Note that for arousal, this result approaches the agreement of groups of five human judges with each other (0.88). For valence, however, it is substantially below the human agreement of 0.91.

Jan van Santen evaluated the hypothesis, which was already confirmed for perceptual ratings, that cross-modal correlation is lower for valence in HFA than in TD. He substituted the automatically predicted scores in the place of the perceptual ratings. We trained the models on 80% of all stimuli, predicted the ratings for the remaining 20% to avoid overfitting the data, and repeated this process five times to have predictions for all utterances. Unfortunately, the precision of the

models did not prove to be enough for replicating the above finding using our predictions.

8.3.2 Predicting perceptual ratings for the CSLU ERPA ADOS Corpus

We evaluated the performance of the regression model by calculating the correlation between the gold standard and the predicted scores. We could not predict ratings for some utterances whose features were invalid (probably due to these containing too few voiced frames), and excluded those utterances from both the perceptual and the predicted rating sets. This way 4292 of the original 4521 remained. The correlation between the actual and the predicted perceptual ratings were: 0.78 for arousal, 0.40 for valence, 0.30 for prosodic atypicality, and 0.23 for prosody–content incongruity. The performance for emotions is nearly identical to that for the CSLU Cross-Modal Corpus (see Section 8.3.1), but there are important differences: The feature set is different in that we use per-subject features as well, calculated from a substantial amount of speech for each subject. The inter-rater agreement for this data set was however substantially worse, which must have a negative effect on the performance on the test-set.

We compared both the perceptual ratings and the predictions across diagnostic groups matched on age only to have more subjects and so possibly better power. Monte Carlo tests did not give significant results for either the perceptual or the predicted ratings for either comparison. We calculated Cohen's d effect sizes between the group scores (see Table 8.2). Based on a Shapiro–Wilk test, the distribution of the ratings was not normal, therefore we report bootstrap estimates using the `bootES` R package (Gerlanc and Kirby, 2015). We can see that for arousal and prosodic atypicality, the effect sizes of the predicted ratings are comparable to that of the perceptual ratings.

Table 8.2: Effect sizes for the gold standard and the predicted perceptual ratings for age-matched diagnostic group pairs from the CSLU ERPA ADOS Corpus. We predicted the ratings using an SVR model with 60 per-speaker and per-utterance statistical moments of F0 and intensity.

	ALN–TD perceptual	ALN–TD predicted	HFA–TD perceptual	HFA–TD predicted
arousal	.25 (.16 – .34)	.27 (.17 – .35)	.19 (.11 – .26)	.29 (.21 – .35)
valence	-.07 (-.16 – .03)	.26 (.16 – .34)	-.04 (-.12 – .03)	.15 (.08 – .22)
prosody atypical	.15 (.05 – .24)	.21. (.12 – .30)	.15 (.07 – .22)	.20 (.12 – .27)
prosody incongruous	.09 (.00 – .18)	06. (-.03 – .15)	.06 (-.02 – .14)	-.01 (-.08 – .07)

8.4 Discussion of the Prediction Results

Our main goal with this work was to see if perceptual ratings predicted using acoustic-prosodic features are suitable for differentiating speakers with HFA from those with TD the same way that the original perceptual ratings are. For example, Jan van Santen showed based on an acted emotional speech corpus that the correlation between valence measures among different modalities, such as speech and text, are significantly lower for speakers with HFA than those with TD. We also showed in Chapter 7 of this work on a spontaneous speech corpus that the arousal ratings were higher in the HFA group than in the TD group, just as the prosodic atypicality ratings for certain stimulus sets.

In this work, we evaluated acoustic-prosodic features for predicting emotions and prosodic atypicality from speech. The features were robust and non-robust moments of F0 and intensity and functional PCA features. There is a substantial amount of literature on recognizing emotions from speech, with a growing number dealing with the dimensional representation (valence, arousal, dominance, control; see e.g. Schuller and Devillers, 2010; Truong et al., 2012; Räsänen and Pohjalainen, 2013; Bone et al., 2014b; Youssefi, 2015). Most of the existing methods work with thousands of features, including spectral, F0, and intensity features, as well as textual features from automatic speech recognition systems. Working with a relatively small number of easy-to-extract features, in our case 30 to 54, has obvious advantages if their performances are comparable.

Our results indicate that acoustic-prosodic features alone may not be sufficient for predicting valence with enough precision. Using our predicted ratings, the cross-modal valence correlation

difference could not be confirmed, but there are lessons to be learned from our experiments. While the raters were able to recognize valence to some extent even without knowing the text, although much worse than from text alone, statistical features of F0 and intensity gave mediocre prediction performance. Adding features that reflect the intonation curve shapes, namely functional PCA features, increased the performance significantly. This indicates that the intonation curve shapes play a role in expressing valence.

Regarding atypicality, we expected the performance to be relatively low, as they indeed were, since it has considerably lower inter-rater agreement in our data set than either of the emotional ratings (0.49). The predicted atypicality had a relatively low correlation with the actual values (0.30), yet the effect size between the diagnostic groups was comparable. The reason is probably that we used per-speaker features as well, which were useful in predicting the overall atypicality per child, and thus the difference between the diagnostic groups was reflected in the outcome even though the accuracy for the individual utterances was low. A similar phenomenon is observable for most of the other ratings as well, which points to the utility of using per-speaker features calculated from all speech available for the speaker even when the number of utterances for supervised training is limited. Note that incongruency had an even worse inter-rater agreement than atypicality (0.45), and even more importantly, it is a quality of the relationship between prosody and content. Since we only used prosodic features, it is not surprising that the performance is low for incongruency.

There is not much previous work on predicting prosodic atypicality from speech. We know of just one by Bone et al. (2015): They conducted a study on perceptual ratings for a corpus of story retellings (Grossman et al., 2013). Children with ASD and TD listened to a story read by an actor, and then retold it by reading out the text. They used several feature sets for predicting the ratings, including rate/rhythm, voice quality, and intonation, besides comparisons to the actor's speech and measuring deviations from the text. They were able to predict "awkwardness", an equivalent of what we called prosodic atypicality, with 0.56 correlation between the actual and the predicted rating (the inter-rater agreement was 0.57), mostly using rate and rhythm features. Some fundamental differences between this study and ours are that they analyzed read speech with (nearly) identical content for each child. They calculated some features by making use of the latter fact: Their transcript-matching features reflect deviations from the prescribed text, and their exemplar-based intonation and stress features compare a stylized version of an individual's intonation curves to one derived from the best instances of the other speakers for the same utterance. We do not know which, if any, of the children with ASD also had language impairment, but we know that the group as a whole differed significantly on performance IQ and receptive vocabulary size. The rate and rhythm features that they found to be most predictive of the atypicality ratings

on this data set were: increased pausing, more variable syllable durations, less variable syllable intensity, and slower speaking rate. Based on the above, we cannot exclude the possibility that differences in the intelligence and reading ability of the children may have played a substantial role in the kind atypicality displayed by them during this task. It is not clear that the features that captured these differences would perform similarly well on spontaneous speech corpora; it is an interesting research direction to evaluate that.

For arousal, the performance of the feature set comprising statistical moments of F0 and intensity was not far from the ceiling determined by the agreement between the raters (0.83 vs. 0.88 on the cross-modal and 0.78 vs. 0.85 for the ERPA ADOS corpus). Note that the performance for the two corpora were quite similar, although the kind of stimulus differed: we worked both with speech from a story-retelling task and with spontaneous speech. The effect size for the predicted ratings between the diagnostic groups is also comparable to that of the perceptual ratings. This finding indicates that an accurate characterization of the F0 and intensity histograms through robust statistics and higher-order moments goes a long way toward recognizing arousal from speech, reaching performances comparable to the agreement of the human ratings.

The performance of the models depends substantially on the amount and quality of the training data. We worked with moderate data sizes: 835 utterances with fairly reliable ratings in one case, and 4292 utterances with lower inter-rater agreement in the other case. Increasing the training set size or a reduction in the standard error of the ratings through collecting more ratings is due to increase the performance. This is especially true for valence, where we concluded that the intonational curve shapes must be relevant as well. Since these have a much higher degree of freedom than the statistical moments, having more examples to learn from must be especially important for predicting valence.

8.5 Summary for Predicting Perceptual Ratings

We evaluated the suitability of F0 and intonation features for predicting human perceptual judgments with sufficient precision to give these automated scores diagnostic relevance. The ratings studied were emotional arousal and valence as well as prosodic atypicality ratings. Robust statistics including those of higher-order moments enabled us to predict arousal with a reliability comparable to that of human judges. Performance for prosodic atypicality was low using these features, however. Adding functional principal component features for characterizing the intonation curve shapes improved the performance for valence, but it was still below the acceptable level. Evidence suggests that precise prediction of valence at the utterance level requires textual as well as

spectral features and presumably more utterances for training. Notwithstanding, when including per-subject statistical features of the intonation curves based on a substantial amount of speech for each individual, the effect sizes for diagnostic group comparisons were on a par with those of the human ratings.

Chapter 9

Conclusions

The topic of this work was speech prosody in Autism Spectrum Disorders (ASD). Both clinical experience reflected in the earliest descriptions and formal research studies on expressive and receptive prosody have shown that there are differences in the speech of children with ASD compared to those of typical development. How their speech differs and what aspects are different has not been conclusively decided yet. Part of the reason must be the heterogeneity of autism: While some individuals never become verbal, which is thankfully not the usual case, others excel not just in certain subject areas but also in speech and writing. But even when their speech is fully functional, it may still sound atypical to the listeners. Characterizing this difference both acoustically and perceptually can have multiple uses: It can advance our understanding of what constitutes the main challenges for these people and can direct prosodic remediation. Characterization of the autism phenotype may help differential diagnosis, and possibly in identifying subgroups in the autistic population, which in turn can help in genetic research. Computational features can provide outcome measures for treatment research. Screening procedures can benefit from automated speech analyses methods, which are non-intrusive and can be applied in a wide range of contexts, including remote communication.

Our main goals were to examine speech prosody in the context of autism compared to typical development and language impairment from a range of aspects: First, whether there are acoustic-prosodic differences that are consistently present between the groups being studied. Second, whether naive listeners can reliably identify atypicalities based on brief instances of speech. Third, how prosodic differences are related to the contents of the speech. Fourth, we wanted to do the above for spontaneous speech utterances. Most previous research dealt with answers to a specialized target task designed to tap into abilities that may be lacking or deficient in autism. Finding autism-specific differences in spontaneous speech can be more challenging, as we do not know in advance what type of differences to expect and where. We did not concentrate on the functional aspects of prosody, but rather on whether it sounds typical or not.

We worked with a corpus of conversational speech from children recorded during the administrations of an autism diagnostic instrument, the ADOS. The subjects were children with High-Functioning Autism (HFA) with or without a comorbid language impairment and children with Specific Language Impairment (SLI) or Typical Development (TD) as controls. While the criterion of including only children with HFA inevitably reduces the heterogeneity compared to that observed in the general autistic community, excluding for example those that are non-verbal, it is arguably the population where research on atypical prosody has the most relevance because it may not be immediately obvious from the fluency, content, or grammaticality of their speech what issues they are facing.

We analyzed speech recordings using standard and novel approaches, making use of both objective measures and perceptual ratings of speech prosody. Naturally our research questions and hypotheses were influenced by the results of previous studies, but we did not restrict our search to the areas indicated by those findings in seeking for diagnostically relevant features. In Chapter 5, we laid one of the foundations of this work by selecting the matched subjects for the group comparisons using complex criteria and a systematic approach to matching. In Chapter 6, we studied acoustic features of speech prosody through the use of three sets of acoustic features: First, robust statistics and higher-order moments of the fundamental frequency and intensity curves; second, speaker-specific intonation model parameters for the Simplified Linear Alignment model estimated for each children; third, the difference in the contribution of various intonation curves that their speech prosody can be decomposed into, using functional principal component analysis. We devised a novel approach for deriving the speaker intonation characteristics that made use of artificial training data synthesized with a text-to-speech system. In Chapter 7, we characterized prosodic differences in HFA compared to TD and SLI through the eyes of naive listeners. It involved the selection of two types of stimulus sets matched between the groups on complex criteria, the design of an interface for collecting subjective ratings for the speech and text modalities, and the assessment and analysis of these new research data sets. We compared the ratings between the groups by controlling for the effect of content features, including the current activity, the utterance type, and numeric utterance characteristics. In Chapter 8, we went further with the perceptual ratings data and trained machine learning algorithms to estimate these for new speech utterances. Using acoustic-prosodic features only, we were able to predict speech arousal with an accuracy that is comparable to the agreement between human judges.

We found or confirmed several significant differences between the diagnostic groups; see Table 9.1. This includes differences in the statistical properties of F0, especially higher spread of the F0 values in HFA than in TD. This refers to global properties of all pitch samples of a subject, which of course comes about as the result of utterance-level differences: For speakers with HFA, the spread of F0 values within the utterances is on average higher than in TD, and the amount of this spread is also more variable across the utterances. The atypical groups were all very similar in these aspects and did not differ from each other significantly after FDR correction on any statistical feature of F0. The diagnostic group also had a significant effect on the speaking rate: The language impaired groups (those with either SLI or ALI) had a significantly lower average speaking rate than the typically developing group. This difference did not reach significance for the ALN–TD comparison.

The properties of the characteristic phrase curve type also differed significantly: The pitch height at the start and middle of the phrases was generally higher in HFA than in TD. We found other evidence regarding differences in the intonation curve types by decomposing the intonation curves into component curves that explain most of the variance among thousands of utterances with similar lengths. The contribution of some of these component curves was also significantly different between the groups. Though this finding is harder to interpret, the numeric features derived from the process proved to be useful in multiple tasks we dealt with, including automatic emotion recognition. In none of the above features did we find significant differences between children with autism and those with language impairment without autism.

Regarding the perceptual ratings, even though the raters seem to have varied widely in their competence and their agreement was low on some of the questions, their combined ratings revealed some statistically significant differences between the groups. For example, we found that arousal in HFA was generally higher than in TD, especially for utterances that were matched on expected prosody. For other stimuli that were prosodically diverse, the raters perceived the speech of children with HFA as more atypical more often. Based on earlier research findings and on our analyses, it seems likely that children with HFA both experience more arousal during the interactions and that listeners interpret some of their prosodic peculiarities as signs of higher emotional arousal. We did not find significant differences between children with HFA and SLI on the perceptual ratings either.

It is important to note here that the agreement between the ratings of the naive judges we worked with was not substantial for prosodic ratings. We strived to make the instructions and the task interface clear and easy-to-use, and incorporated feedback both from colleagues and from raters in the process. We also increased the number of ratings per utterance used for deriving aggregated scores from five to ten, and this increased the agreement substantially. Even so, the agreement was high for emotional arousal and valence only, moderate for unusuality of content (unusual meaning and unusual words), and moderate to low for prosodic atypicality and prosody–content incongruency. Providing more direction to raters may help in this regard. As it is, since they were blind to the purpose of the tasks, different judges might have had differing ideas about what counts as atypical. Some may have looked for pathological differences only, others may have looked for anything that deviated from the ideal.

One of our unexpected findings, confirmed again and again using the various approaches utilized, was that speech prosody in language impairment differed from that of typical development similarly to what has been observed in autism. This came somewhat as a surprise because this aspect of language impairment is rarely mentioned in the literature. For example Peppé et al.

(2011) stated that "atypical expressive prosody is not usually observed as a feature of specific language impairment." Conversely Chown (2012), while also quoting Peppé, mentions that McCann et al. (2007) "have drawn attention to prosodic similarities between autism and SLI." He goes on to say that the prosodic impairments in autism may be due to language impairments, just as Peppé et al. (2011) also says that in relationship to prosodic impairments, the overlap between the ASD and SLI diagnoses may be relevant. These statements would implicate that autism without language impairment should not be associated with prosodic impairment. While this may be true for functional aspects of prosody, our findings show that the group of autistic children without language impairment differs from the typical group significantly in their prosodic expression. In summary, the prosodic characteristics of the SLI group did not differ significantly from that of the HFA group despite that they differed very markedly on features associated with autism. It is true that in some of our analyses, we matched the utterance content features very closely between the groups, as we wanted to compare prosody irrespective of language characteristics. This may have eliminated some differences between the HFA and the SLI groups that would show up when examining the whole range of utterances. In this work, however, we wanted to identify prosodic differences between HFA and SLI that cannot be explained by language abilities, but we were not able to find such difference. It is important to note that this was the case even though all clinicians in the consensus meetings agreed that these children with SLI did not display any one of the three DSM-IV criteria for autism.

Table 9.1: Summary of main findings on prosody in autism. When not otherwise specified, the group with autism is compared to the typically developing controls.

Topic	Finding	Section
Acoustic features	Different shape of average F0 histograms confirmed	6.3.1
	Significant difference in several statistical moments of F0 confirmed	6.3.1
	Significantly lower speaking rate for those with ALI or SLI	6.3.1
	Phrase start and inflection point intonation model parameters significantly higher	6.3.2
	Certain F0 Functional PCA coefficients significantly different	6.3.3
Perceptual ratings	Greater emotional arousal after controlling for content features	7.3.2, 7.3.4
	Somewhat higher perceived prosodic atypicality only when content features are not matched	7.3.4
	A trend to have different atypical prosodic aspects compared to typically developing children	7.3.5
Predicting emotional ratings	We can predict emotional arousal with a reliability comparable to human judges using statistical features of F0 and intensity	8.3.1, 8.3.2
Prosody in Language Impairment	We did not find significant differences between the groups with HFA and SLI after controlling for content features	all of the above

While we made efforts to ensure that our results are reliable, most of these are the outcomes of analyses on just one corpus. Replicating them on other corpora and possibly other control groups is necessary to increase their trustworthiness. One candidate is the corpus of ADOS sessions recorded at the Fair Neuroimaging Laboratory: This laboratory, headed by Damien Fair, professor of behavioral neuroscience and psychiatry, collect and analyze data for children with autism as well as ADHD, and TD controls. The main focus is on fMRI studies, but the data include speech and cognitive measurements as well.

Limitations of this study include that we mostly dealt with the speech intonation, only touching on other aspects of prosody, namely the variations of rhythm and loudness. Moreover, we dealt with voice quality only through perceptual ratings and found no difference in this regard — but our listeners had a low agreement on this aspect of speech. So the lack or scarcity of findings in these other areas does not mean that the groups do not differ significantly there. The heteogeneity

of the prosody of individuals with ASD may also make it harder to come up with generalizations as some individuals may lack such symptoms and for others, the manifestation may be very diverse.

As with any other work, every answered question raises more questions and every finding has many alternative explanations that cannot all be ruled out in the scope of a finite-length study. There are too many interesting and possibly important research directions to be listed here, both to examine so far unevaluated aspects of prosody and to dispel doubts regarding the reliability of existing findings. We will mention just a few.

Our results hinge on the quality of the prosodic features, especially that of the F0 curves, but so far no F0 detector is a match for the reliability of human perception. It may be worth ensuring the trustworthiness of our intonation curves using human effort. In a preliminary experiment, we played the original waveforms and their vocoded versions using the result of several F0 detectors to human subjects, asking them to point out any differences between their pitch. The F0 tracking method we relied on in this work got the best scores; yet it is not perfect either. By going one step further, this approach can be extended to be the basis of a semi-automatic correction of the erroneous curves.

In this work, our unit of analysis was the individual utterance, corresponding to the everyday concept of a sentence. We touched on interactional aspects of speech, yet there is much more that would be worth investigating, especially the interaction between the child and his or her communication partner (in our corpus, the examiner; see e.g. Levitan et al., 2011; Levitan and Hirschberg, 2011; Bone et al., 2014a; Hopkins et al., 2016).

Echolalia is another aspect of autism that has been frequently described from the beginnings (Kanner, 1943). It is the (immediate or delayed) repetition of the conversation partner's words by the child, and it can involve the imitation of the prosody as well or modifying it for communicative purposes (Paccia and Curcio, 1982). van Santen et al. (2013) have quantified their occurrences in the transcripts, but to our knowledge, the same has not been done for the prosody of the echoed utterances.

We addressed some questions by studying a corpus of spontaneous speech utterances from children with HFA, SLI, and TD. Beyond the limitation due to using only one corpus, and that with a restricted number of subjects in the corpus, our results cannot apply to non-verbal or minimally verbal children with ASD, and may not apply to those with comorbid intellectual disability as these groups were not represented in our corpora. In possession of data from subjects with a wide range of intellectual abilities, one would be able to examine the dependence of various prosodic features on IQ.

We looked at a cross-section of the prosodic behavior of the children, but longitudinal studies have much to contribute to our understanding. They can show for example which aspects of the children's speech develop similarly and where children with autism may need additional help. Such data gives insight into the developmental trajectory of prosodic features with age: whether the prosody of children with ASD tends to grow more typical with age and where certain types of differences between the diagnostic categories emerge. This in turn can show which features are more useful for screening and characterization of prosody at different ages. (Note that some of our features, specifically the statistical features of F0, seem to be applicable not just to speech, but also to preverbal vocalizations.) Having multiple measurements per child can also help to eliminate spurious findings on the relationship between prosodic features and other subject characteristics. A corpus that seems suitable for longitudinal research is for example the ADOS Corpus of the University of Washington.

In the past decades, the number of acoustic analyses of prosody has increased substantially, probably both to the growing interest and concern in society about this disorder as well as the availability of the technology and data that are required for this type of study. To avoid a "replication crisis" in this area of research, the findings of these studies need to be replicated by other research groups on other corpora. We hope that others will take interest in replicating our findings as well.

Chapter 10

Vision

Having seen the scientific contributions of this work, our next logical questions are: What can we do with these results to make them useful in practice? And where does this line of research lead over the course of years if someone is going to pursue it? We are going to deal with these questions in this order below.

Possible applications of our knowledge of and automated techniques for analyzing speech prosody include following:

1. Automated assessment or screening: Screening for developmental disorders currently requires trained clinical personnel to administer diagnostic instruments, which is time-consuming and expensive. Computational methods have the potential to provide such assessments unobtrusively during office visits, or potentially even remotely, at a low cost. This is especially the case if they are based on spontaneous speech instead of structured tasks.

2. Differential Diagnosis: Speech abnormalities, including atypical prosody, have the potential utility to identify some medical conditions. However, since such issues are present in multiple disorders, it is of interest to delineate them from each other, if indeed there are phenotypic differences on this level. As we have seen earlier, making this level of distinction (beyond the typical–atypical differentiation) is not trivial, and may not even be possible, but is certainly of interest.

3. Prosodic remediation: Computational methods for assessing the quality of speech and giving feedback can aid such interventions by increasing our understanding of potential deficits and creating applications for training. Potential targets of prosodic remediation include not just the functional aspects, but also the prosodic forms. Getting the latter right can be important for social acceptance of the individual.

4. Outcome measures for evaluating treatments in a research setting: Quantitative assessment of the current level of an individual's prosodic ability can help to assess the effectiveness of interventions by providing objective measures at different time points during the process.

5. Distinguishing prosodic subgroups within ASD: The autistic phenotype is known to be very heterogeneous. Characterizing this heterogeneity regarding speech prosody and identifying prosodic subgroups can help for example genetic research.

Our results are twofold: First, we have identified qualitative differences between the speech of children with autism and those with typical development, and perceptual ratings that differentiate between the groups. We presume that such understanding can help professionals who train individuals on the spectrum to improve aspects of their prosody that are most deviant. It can also

contribute to building prosodic models that best characterize the facets of prosody that differentiate autistic speech from typical speech. Second, we have come up with feature sets and methods to predict these perceptual ratings, and have created sets of acoustic-prosodic features that can differentiate between autistic and typical speech to some extent. These are applicable in systems for automatic assessment of prosody. Obviously, both types of results can help in creating the applications outlined above.

There are several ways this line of research can be continued. One way is to experiment with the parameters of the techniques to improve them. It is immediately obvious that no algorithm or other computational technique is perfect and there are always some ad-hoc decisions whose alternatives can be explored in the hope of getting better results. Another way is to work toward our goals by adding more techniques to our repertoire.

Possible future research involves the following:

- Covering aspects of the interaction between the child and his or her communication partner.
- *Echolalia*, the (immediate or delayed) repetition of the conversation partner's words by the child, can involve the imitation of the prosody as well or modifying it for communicative purposes (Paccia and Curcio, 1982). Van Santen et al. (2013) have quantified their occurrences in the transcripts, but to our knowledge, the same has not been done for the prosody of the echoed utterances.
- Replicating findings of other research groups on our corpora.
- Replicating our findings on other corpora, possibly involving longitudinal data and other control groups (such as ADHD or ID)
- Expanding the scope of our research to non-verbal and minimally verbal children, by analyzing non-speech vocalizations and cry acoustics.

Research on these topics needs to take into account important subject characteristics that can have a substantial effect on how we need to approach our questions. The most important ones seem to be age, intellectual ability, and the heterogeneity of the autistic population. The development of cognitive abilities is non-linear with age, which is likely to be the case for prosodic abilities as well. The concept of heterogeneity is also very relevant because autism is a *spectrum disorder*, which means that, despite commonalities that are shared by all subjects, its symptoms can be present in highly varying degrees and forms. For example, individuals with autism exhibit a wide range of intellectual ability from intellectual disability to savant skills. Earlier research has shown that

speech atypicalities occur in a large percentage of autistic subjects irrespective of their IQ, yet IQ has a substantial effect on the outcome (Matson and Shoemaker, 2009). Children at different cognitive levels may differ in their prosodic phenotype, may need to be assessed differently, and may need differing kinds and amounts of treatment.

Our methods can in theory be used for other languages for which we have similar corpora. One exception is the estimation of the speaker-specific intonation curve parameters, which also requires a TTS system that implements the SLAM intonation model (see Section 3.2.3). The set of content features may also need to be altered depending on the language. For example, it seems better to use the number of syllables or number of morphemes instead of the number of words for synthetic languages (those with a high number of morphemes per word).

Technical obstacles to assessing prosody in spontaneous speech include the lack of reliable F0 detection requiring low computational resources and of reliable Automatic Speech Recognition (ASR). The latter is an issue because content affects prosody, so ideally we need to be able to recognize the content to be able to control for its effect. These may be addressed suitably by using F0 features that are robust to tracker errors and content. The quality of ASR systems is also continually improving.

Fruitful research in these areas can lead to robust feature sets that may be able to approach the ideal solution. An ideal assessment tool would have normative data from a web-scale speech corpus that enables it to model typical speech prosody to the extent that it can quantify the allowable amount and type of deviations from the ideal as well. It would also be able to screen obvious cases, those with a high level or amount of atypical speech, with an essentially perfect recall. Moreover, it would be able to identify even rare and subtle signs that may indicate the presence of a disorder in order to refer the patient for further evaluation. Having models trained on large corpora with subjects from the whole range of the spectrum from diverse atypical populations would enable us to use subtle prosodic cues (and possibly even language features from an ASR system) for differentiating between various disorders. With the advancement of machine learning and the availability of corpora of increasing sizes, this ideal will become increasingly realistic.

Appendix A

Perceptual Ratings of Prosody: Supplementary Materials

A.1 Simple Task Interfaces (S) for Collecting Perceptual Ratings

A.1.1 Text rating task (TEXT)

Rate Emotional Charge and Meaningfulness of Children's Sentences

Instructions (the same for each HIT):

Please take this HIT only if you are a native speaker of English.

You can see ten sentences or sentence fragments from children, together with the approximate age of the child. We left out sentence final punctuation; you decide for yourself if it is a statement, an exclamation, or a question.

First select *how negative or positive* the child probably felt when saying it.

Second, select *how calm or excited* the child probably felt. Choose "I don't know" if and only if all emotions seem equally likely, otherwise choose the most likely one. If multiple emotions are present, please choose the dominant one.

Third, tell us if *any of the words individually seems unusual* for any reason (the word is unusual for a child of this age, the word does not exist, etc.).

Fourth, tell us if *the sentence as a whole has an unusual meaning* (including the case that it does not make sense).

For example: "I like her". You would probably rate this as "somewhat positive" on the emotional positivity scale, "neither calm nor excited" (neutral) or "somewhat excited" on the activation scale, plus you would choose that it has no unusual words and no unusual meaning.

There are usually no right or wrong answers, we just want to know your opinion. If you do your best to follow the instructions, we will definitely pay you. But if you provably do not follow the instructions in all cases (e.g. if you choose a random answer in certain cases), then we may not pay you, and may even block you.

Thank you for your work!

Questions

Sentence $\langle N \rangle$ **:** $\langle$**sentence text**$\rangle$ $\langle n \rangle$-year-old boy/girl

How do you think the child saying this felt?

( ) Very negative ( ) Somewhat negative ( ) Neither negative nor positive ( ) Somewhat positive ( ) Very positive ( ) I don't know

( ) Very calm ( ) Somewhat calm ( ) Neither calm nor excited ( ) Somewhat excited ( ) Very excited ( ) I don't know

Does the sentence contain an **unusual word or words?**

( ) No ( ) Yes

Does the sentence have an **unusual meaning overall?**

( ) No ( ) Yes

A.1.2 Speech rating task (SPEECH)

Rate Emotional Charge and Meaningfulness of Children's Sentences

Instructions (the same for each HIT):

Please take this HIT only if you are a native speaker of English and have a good ear for speech. Please do the HITs in a quiet environment, and use a headset if possible. Thank you!

You are going to listen to ten sentences or sentence fragments from children. **Please listen** to each sentence by pressing the play button, **then answer** some questions about it, **concentrating on *how*** the child said it (that is the intonation, rhythm, loudness, speed, voice quality, and similar), *not what* the child said, and disregarding any pronunciation errors as well. You can see the approximate age of the child to the right of the player.

First, select how negative or positive the child probably felt when saying it. Second, select how calm or excited the child probably felt. Choose "I don't know" if and only if all emotions seem equally likely, otherwise choose the most likely one. If multiple emotions are present, please choose the dominant one. Third, tell us if *how* he or she uttered the sentence is typical or unusual for a child in this age range. In other words, choose "Somewhat / Very unusual" if the child uses an intonation, rhythm, etc. that children of this age normally do not use for any sentence.

Finally, please consider if *how* the child spoke and *what* the child said agree with each other, or if they are in some way incompatible. In other words, choose "They mismatch somewhat / completely" if how the child spoke is strange for this content (although it may be suitable for some other specific content). For example, the child speaks about something distressing in a casual way, or the emphasis is not on the words where it should be, etc.

Some sentences contain so called "mazes": filled pauses (e.g. "uh", "um"), false starts, and

repetitions and revisions of words. We marked these in the textual transcript (shown below the player) by putting parentheses around them. For example: "(Uh Ca Can I) I'm going to stand on it". The presence of mazes is normal in everyday speech, and thus their presence in itself should not be considered unusual are atypical, unless of course the child uses them in an unusual or atypical way.

There are usually no right or wrong answers, we just want to know your opinion. If you do your best to follow the instructions, we will definitely pay you. But if you provably do not follow the instructions in all cases (e.g. if you choose a random answer in certain cases), then we may not pay you, and may even block you.

Thank you for your work!

Examples of how the sentence can sound unusual: the wrong words are emphasized; it sounds monotonous or singsong; the pitch is too low / too high / too flat / too varied; the location / length / frequency of pauses is unusual; the speed is too slow / too fast / too varied; the child spoke too softly / too effortfully / with uneven loudness; the voice is very tense / very hoarse / too nasalized; and anything else that makes the sentence sound strange or unusual.

Questions

Sentence ⟨N⟩ : ⟨waveform player⟩ ⟨n⟩-year-old boy/girl

⟨sentence text⟩

How do you think the child saying this felt?

Please concentrate on how the child said it, not what the child said.

( ) Very negative ( ) Somewhat negative ( ) Neither negative nor positive ( ) Somewhat positive ( ) Very positive ( ) I don't know

( ) Very calm ( ) Somewhat calm ( ) Neither calm nor excited ( ) Somewhat excited ( ) Very excited ( ) I don't know

Did the child say this sentence in **a typical or an unusual, strange way**?

Please pay attention to how he or she said it, not what was said, disregarding any articulation errors as well.

( ) Typical ( ) Somewhat unusual ( ) Very unusual

Do the "what" (the sentence content) **and the "how"** (the way it is said) **match**?

( ) They match well ( ) They mismatch somewhat ( ) They mismatch completely

A.1.3 Delexicalized speech rating task (DELEX)

Rate Intonation and Emotions in Children's Sentences from Blurred Speech

Instructions (the same for each HIT):

Please take this HIT only if you are a native speaker of English and have a good ear for speech. Please do the HITs in a quiet environment, and use a headset if possible. Thank you!

Please listen to each short, blurred, incomprehensible sentence below by pressing the play button, **then answer** some questions about it, **concentrating on the way** the child spoke. You can see the age and gender of the child to the right of the player.

First select how negative or positive the child probably felt when speaking. Second, select how calm or excited the child probably felt. Choose "I don't know" if and only if all emotions seem equally likely, otherwise choose the most likely one. Third, tell us if the way he or she spoke (that is the intonation, speed, loudness, and similar) is typical or unusual for a child of this age.

There are usually no right or wrong answers, we just want to know your opinion. But there are enough clear-cut cases for us to know if you are really paying attention to what you are doing.

Thank you!

Examples of how the sentence can sound unusual: it sounds monotonous or singsong; the pitch is too low / too high / too flat / too varied; the speed is too slow / too fast / too varied; the child spoke too softly / too effortfully / with uneven loudness; and anything else that makes the sentence sound strange or unusual.

Some helpful tips: You can bring all questions for a sentence into view by clicking on the "Sentence" link. You can submit the HIT by pressing ENTER.

Questions

Sentence $\langle N \rangle$: $\langle$waveform player$\rangle$ $\langle n \rangle$-year-old boy/girl

How do you think the child speaking felt?

( ) Very negative ( ) Somewhat negative ( ) Neither negative nor positive ( ) Somewhat positive ( ) Very positive ( ) I don't know

( ) Very calm ( ) Somewhat calm ( ) Neither calm nor excited ( ) Somewhat excited ( ) Very excited ( ) I don't know

Did the child speak in **a typical or an unusual, strange way**?

Please pay attention to how he or she spoke, disregarding that we rendered the contents unintelligible.

( ) Typical ( ) Somewhat unusual ( ) Very unusual

A.1.4 Speech aspect rating task (SPEECH ASPECTS)

Identify Unusual Aspects of the Intonation of Children's Speech

Instructions (the same for each HIT):

Please take this HIT only if you are a native speaker of English and have a good ear for speech. Please do the HITs in a quiet environment, and use a headset if possible. Thank you!

Please listen to the sentence below by pressing the play button, **then answer** some questions about it, **concentrating on *how*** the child said it (that is the intonation, rhythm, loudness, speed, voice quality, and similar), *not what* the child said, and disregarding any pronunciation errors as well. You can see the age and gender of the child to the right of the player.

Please indicate if different aspects of the way the child uttered the sentence are typical or unusual for a child of this age, and in exactly what ways the sentence sounds unusual or strange.

There are usually no right or wrong answers, we just want to know your opinion. But there are enough clear-cut cases for us to know if you are really paying attention to what you are doing.

Thank you!

Questions

Sentence: ⟨waveform player⟩ ⟨n⟩-year-old boy/girl

It **sounds monotonous**. ( ) No ( ) Yes

It **sounds singsong**. ( ) No ( ) Yes

The wrong words are emphasized (bad stress placement). ( ) No ( ) Yes

The **location, length, or frequency of pauses is atypical**. ( ) No ( ) Yes

The **pitch is too low**. ( ) No ( ) Yes

The **pitch is too high**. ( ) No ( ) Yes
The **pitch is too flat**. ( ) No ( ) Yes
The **pitch is too varied**. ( ) No ( ) Yes
The **pitch is atypical in some other way**. ( ) No ( ) Yes

The **speed is overall too slow**. ( ) No ( ) Yes
The **speed is overall too fast**. ( ) No ( ) Yes
Some parts are much faster than other parts. ( ) No ( ) Yes
The **speed is atypical in some other way**. ( ) No ( ) Yes

The child **spoke too softly**. ( ) No ( ) Yes
The child **spoke too effortfully**. ( ) No ( ) Yes
Some parts are much louder than other parts. ( ) No ( ) Yes
The **loudness is atypical in some other way**. ( ) No ( ) Yes

The voice is **very tense**. ( ) No ( ) Yes
The voice is **very hoarse**. ( ) No ( ) Yes
The voice is **too nasalized** (hypernasal). ( ) No ( ) Yes
The **voice quality is atypical in some other way**. ( ) No ( ) Yes

A.2 Detailed Task Interfaces (D) for Collecting Perceptual Ratings

A.2.1 Text rating task (TEXT)

Rate Emotional Charge and Meaningfulness of Children's Sentences

Instructions (the same for each HIT):

Please take this HIT only if you are a native speaker of English, and have interacted with young children (between the ages of 4 and 8) a lot. Thank you!

OVERVIEW: You are going to see 25 sentences or sentence fragments from children, together with the gender and approximate age of the child, and answer some questions about them. We

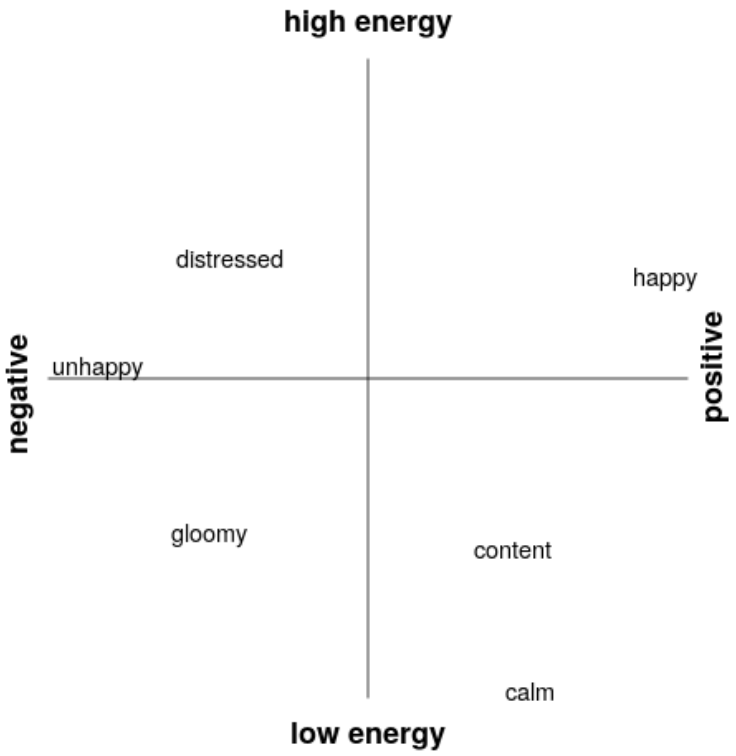


Figure A.1: Examples for some emotions in the arousal–valence plane

left out punctuation from the end of the sentence; you decide for yourself if it is a statement, an exclamation, or a question.

STEPS FOR EACH SENTENCE:

STEP 1: Select how negative or positive and how low or high energy the child probably felt when saying it, by clicking on a location in the coordinate system. See examples for some emotions in the figure.

STEP 2: Indicate how confident you are in your rating.

STEP 3: Tell us which of the words, if any, seems unusual for any reason. This includes saying things in odd or indirect ways (like "blood tubes" instead of "veins"), using words that s/he seems to have invented or made up (like "filpops"), or using words or phrases that sound more grown-up (like "metaphorically speaking", or "a mighty uproar of laughter").

STEP 4: Tell us if the sentence as a whole has an unusual meaning, including the cases that it does not make sense or is about a topic you would not expect from a child of this age.

EXAMPLE: "I like her". You would probably rate this as somewhat positive and somewhat energetic, plus you would choose that it has no unusual words and no unusual meaning.

PLEASE FOLLOW THE INSTRUCTIONS CAREFULLY! There are usually no right or wrong answers, we just want to know your opinion. However, we may not pay you if it is clear that you were not following the instructions (e.g. if you choose a random answer sometimes).

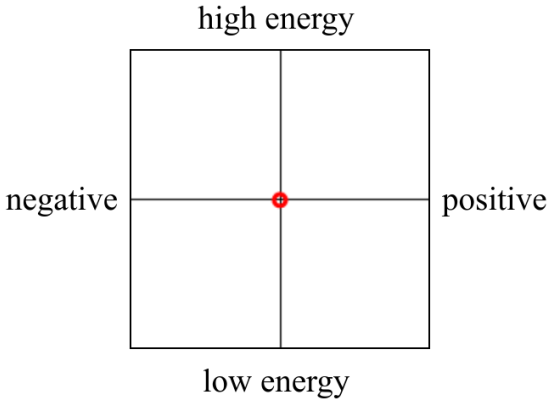
Answers to frequently asked questions: Yes, you can do as many HITs as many are available to you, not just one. If you forget to supply all answers, you will be taken back to the last (hopefully only) missing answer, which will be surrounded by a red box.

Thank you for your work!

Questions

Sentence $\langle N \rangle$ **:** $\langle$**sentence text**$\rangle$ kindergarten/school- boy/girl

How do you think the child saying this felt?



How **confident** are you in your above rating?

( ) Not at all ( ) Somewhat ( ) Very

Which word or words are **unusual** if any? Please click either on "NONE" or on each unusual word.

NONE ⟨ word1 ⟩ ⟨ word2 ⟩ ...

Does the sentence have an **unusual meaning overall?**

( ) Not at all ( ) Somewhat ( ) Very

A.2.2 Speech rating task (SPEECH)

Rate Emotional Charge and Meaningfulness of Children's Sentences

Instructions (the same for each HIT):

Please take this HIT only if you are a native speaker of English, and have interacted with young children (between the ages of 4 and 8) a lot.
Please do the HITs in a quiet environment, and use a headset if possible.
Thank you!

OVERVIEW: You are going to listen to 25 recordings of children speaking. **Please listen** to each utterance by pressing the play button. **Then answer** some questions about it, **concentrating on how** the child said it (that is the intonation, loudness, rhythm, speed, pausing, and voice quality), not what the child said and disregarding pronunciation errors.

STEPS FOR EACH UTTERANCE:

STEP 1: Select how negative or positive and how low or high energy the child probably felt when saying it, by clicking on a location in the coordinate system. See examples for some emotions in the figure.

STEP 2: Indicate how confident you are in your rating.

STEP 3: Tell us if how he or she spoke is typical or unusual for a child in this age range. In other words, choose "Somewhat unusual / Very unusual" if the child speaks in a way that children of this age would not be likely to use. You can see the gender and approximate age of the child to the right of the player. For example:

- The pitch is too low / too high / too flat / too varied.
- The speed is too slow / too fast / too varied.
- The child spoke too softly / too effortfully / with uneven loudness.
- The voice is very tense / very hoarse / too nasalized.
- The length / frequency of pauses is unusual.

STEP 4: Please consider if how the child spoke and what the child said agree with each other, or if they are in some way incompatible. In other words, choose "They mismatch somewhat / substantially" if how the child spoke is strange for this content (even though it may be suitable for some other specific content). For example:

- The child says: "I enjoyed the party so much" in a sad way.
- The wrong words are emphasized.
- The child pauses at unusual places.

The transcript of the recording is displayed below the player. Words that do not contribute to the meaning of the sentence are surrounded in parentheses. They are normal in speech and should not be considered unusual unless the child uses them in an unusual or atypical way.

PLEASE FOLLOW THE INSTRUCTIONS CAREFULLY! There are usually no right or wrong answers, we just want to know your opinion. However, we may not pay you if it is clear that you were not following the instructions (e.g. if you choose a random answer sometimes).

Answers to frequently asked questions: Yes, you can do as many HITs as many are available to you, not just one. If you forget to supply all answers, you will be taken back to the last (hopefully only) missing answer, which will be surrounded by a red box.

Thank you for your work!

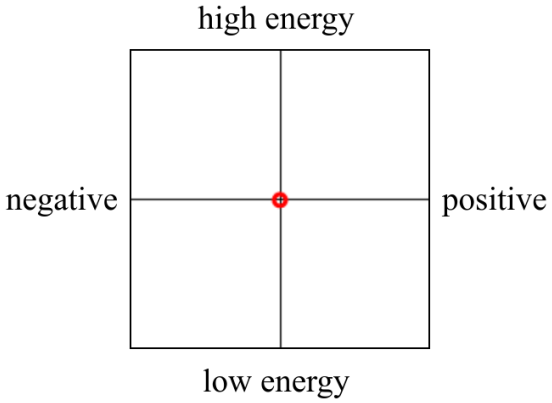
Questions

Utterance ⟨N⟩ : ⟨waveform player⟩ ⟨n⟩-year-old boy/girl

⟨sentence text⟩

How do you think the child saying this felt?

Please concentrate on **how** the child said it, not **what** the child said.



How **confident** are you in your above rating?

( ) Not at all ( ) Somewhat ( ) Very

Did the child say this sentence in a **typical** or an **unusual, strange** way?

Please pay attention to **how** he or she said it, not **what** was said, disregarding mispronunciations, lisps, and stuttering as well.

( ) Typical ( ) Somewhat unusual ( ) Very unusual

Do the **"what"** (the sentence content) **and the "how"** (the way it is said) **match**?

( ) They match well ( ) They mismatch somewhat ( ) They mismatch completely

http://riversidechurch.hu/we-believe/

A.2.3 Speech aspect rating task (SPEECH ASPECTS)

Identify Unusual Aspects of the Intonation of Children's Speech

Instructions (the same for each HIT):

Please take this HIT only if you are a native speaker of English, and have interacted with young children (between the ages of 4 and 8) a lot.
Please do the HITs in a quiet environment, and use a headset if possible.
Thank you!

You are going to listen to 10 recordings from children speaking. **Please listen** to each utterance by pressing the play button. **Then answer** some questions about it, **concentrating on how** the child said it (that is the intonation, loudness, rhythm, speed, pausing, and voice quality), not what the child said and disregarding pronunciation errors.

Please indicate if different aspects of the way the child uttered the sentence are typical or unusual for a child in this age range, and in exactly what ways the utterance sounds unusual or strange. You can see the age and gender of the child to the right of the player.

PLEASE FOLLOW THE INSTRUCTIONS CAREFULLY! There are usually no right or wrong answers, we just want to know your opinion. However, we may not pay you if it is clear that you were not following the instructions (e.g. if you choose a random answer sometimes).

Answers to frequently asked questions: Yes, you can do as many HITs as many are available to you, not just one. If you forget to supply all answers on a page, the missing one(s) will be surrounded by a red box.

Thank you for your work!

Questions

Sentence: ⟨waveform player⟩ ⟨n⟩-year-old boy/girl

The wrong words are emphasized (bad stress placement). Not at all ( ) Somewhat ( ) Completely ( )
Pausing is atypical (e.g. location, length, or frequency). Not at all ( ) Somewhat ( ) Completely ( )

The **pitch is too low**. Not at all ( ) Somewhat ( ) Completely ( )
The **pitch is too high**. Not at all ( ) Somewhat ( ) Completely ( )
The **pitch is too flat**. Not at all ( ) Somewhat ( ) Completely ( )
The **pitch is too varied**. Not at all ( ) Somewhat ( ) Completely ( )
The **pitch is atypical in some other way**. Not at all ( ) Somewhat ( ) Completely ( )

The **speed is overall too slow**. Not at all ( ) Somewhat ( ) Completely ( )
The **speed is overall too fast**. Not at all ( ) Somewhat ( ) Completely ( )
Some parts are much faster than other parts. Not at all ( ) Somewhat ( ) Completely ( )
The **speed is atypical in some other way**. Not at all ( ) Somewhat ( ) Completely ( )

The child **spoke too softly**. Not at all ( ) Somewhat ( ) Completely ( )
The child **spoke too effortfully**. Not at all ( ) Somewhat ( ) Completely ( )
Some parts are much louder than other parts. Not at all ( ) Somewhat ( ) Completely ( )
The **loudness is atypical in some other way**. Not at all ( ) Somewhat ( ) Completely ( )

The voice is **very tense**. Not at all ( ) Somewhat ( ) Completely ( )
The voice is **very hoarse**. Not at all ( ) Somewhat ( ) Completely ( )
The voice is **too nasalized (hypernasal)**. Not at all ( ) Somewhat ( ) Completely ( )
The **voice quality is atypical in some other way**. Not at all ( ) Somewhat ( ) Completely ( )

Appendix B

Code for Reproducible Research

B.1 R Packages

We created part of the infrastructure necessary for creating the analysis pipeline, with a view to ensuring that all our analyses are reproducible. We organized most of the code into R packages, some of which we have already made available within CSLU, OHSU. Below we give an overview of their functionality. We described the packages in detail in the package help, including the public functions and the stored data. (Use `help(package = <package_name)` in an R-session.)

B.1.1 `ldamatch`

This package is for selecting statistically similar research groups by backward selection using various robust algorithms, including a random search, a heuristic based on linear discriminant analysis, multiple heuristics based on the test statistic, and a parallelized exhaustive search.

B.1.2 `GMatcher`

This package serves to create matched tables using the `ldamatch` package. The user needs to create a parameter table with one row for each table containing parameters for the matched table; the package helps the creation of this table as well by generating it from a simple named list. It can create the matched table using various algorithms and store the results from each one. Finally, it can save the matched table with the best characteristics.

B.1.3 `GFeatures`

Provides functions for extracting features from waveforms, F0 curves, and intensity curves. The features include statistical features of time series data, statistics of statistics, and fPCA coefficients.

B.1.4 `GSignif`

Functions for calculating significance values. Currently only Monte Carlo simulations are implemented.

B.1.5 `GPhon`

It contains phoneme set constants for the CMUBET and ARPABET phonetic alphabets, and a function for syllabifying a phoneme sequence, and another one for printing the ones that it syllabifies differently from the syllabification given in a table. See the results of its evaluation briefly in Section 6.2.1.

B.1.6 `GProsodyRatings`

Prepares the rating tasks for comparing the expressive prosody of children with autism, language impairment, or typical development on different types of utterances. Also contains the ratings from two raw and aggregated ratings from two rating collections for four different tasks.

B.1.7 `GAMTRatings`

Provides some utility function for preparing stimuli for Amazon Mechanical Turk experiments and for preprocessing the results.

B.1.8 `GUttChooser`

This package contains the code for choosing ADOS utterances for the perceptual rating experiments using various criteria, such as number of words, if there are mazes, prosodic features, etc.

Bibliography

Emile Aarts and Jan Karel Lenstra. *Local search in combinatorial optimization*. Princeton University Press, 2003.

Sassan Ahmadi and Andreas S. Spanias. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE Transactions on Speech and Audio Processing*, 7(3):333–338, 1999. ISSN 10636676. doi: 10.1109/89.759042.

Cyril Allauzen, Michael Riley, and Johan Schalkwyk. Filters for Efficient Composition of Weighted Finite-State Transducers. In *International Conference on Implementation and Application of Automata*, 2010.

Hiie Allik, Jan-Olov Larsson, and Hans Smedje. Health-related quality of life in parents of school-age children with Asperger Syndrome or High-Functioning Autism. *Health and Quality of Life Outcomes*, 4(1):8, 2006. ISSN 1477-7525. doi: 10.1186/1477-7525-4-1.

Murray Alpert, Enrique R. Pouget, and Raul R. Silva. Reflections of depression in acoustic measures of the patient's speech. *Journal of Affective Disorders*, 66(1):59–69, 2001. ISSN 01650327. doi: 10.1016/S0165-0327(00)00335-9.

American Psychiatric Association and others. DSM-IV-TR: Diagnostic and statistical manual of mental disorders, text revision. *Washington, DC: American Psychiatric Association*, 75, 2000.

American Psychiatric Association and others. *Diagnostic and statistical manual of mental disorders (DSM-5)*. American Psychiatric Publishing, 5 edition, 2013.

American Speech-Language-Hearing Association. Spoken Language Disorders, 2016.

R. Angeleri, F. M. Bosco, M. Zettin, K. Sacco, L. Colle, and B. G. Bara. Communicative impairment in traumatic brain injury: A complete pragmatic assessment. *Brain and Language*, 107 (3):229–245, 2008. ISSN 0093934X. doi: 10.1016/j.bandl.2008.01.002.

Juan Pablo Arias, Carlos Busso, and Nestor Becerra Yoma. Energy and F0 contour modeling with Functional Data Analysis for Emotional Speech Detection. In *Interspeech*, pages 2871–2875, 2013.

Meysam Asgari, Géza Kiss, Jan P. H. van Santen, Izhak Shafran, and Xubo Song. Automatic Measurement of Affective Valence and Arousal in Speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 965–969, 2014. ISBN 9781479928934.

Hans Asperger. Autistic psychopathy in childhood. In Uta Frith, editor, *Autism and Asperger syndrome*, pages 37–92. Cambridge University Press, Cambridge, 1944a. ISBN 0-521-38448-6 (hardcover), 0-521-38608-X (paperback).

Hans Asperger. Die 'Autistischen Psychopathen' im Kindesalter. *Archiv für die Psychiatrie und Nervenkrankheiten*, 117:76–136, 1944b.

J. Ben Atkinson. A greedy look-ahead heuristic for combinatorial optimization: An application to vehicle scheduling with time windows. *The Journal of the Operational Research Society*, 45(6): 673–684, 1994. ISSN 0160-5682. doi: 10.1057/palgrave.jors.2602339.

Kirrie Jane Ballard, Danica Djaja, Joanne Arciuli, Deborah G. H. James, and Jan van Doornd. Developmental Trajectory for Production of Prosody: Lexical Stress Contrastivity in Children Ages 3 to 7 Years and in Adults. *Journal of Speech, Language, and Hearing Research*, 55 (December):1822–1835, 2012. doi: 10.1044/1092-4388(2012/11-0257)1822.

Christiane A. M. Baltaxe. Acoustic characteristics of prosody in autism. *Frontier of Knowledge in Mental Retardation*, pages 223–233, 1981.

Naama Barnea-Goraly, Hower Kwon, Vinod Menon, Stephan Eliez, Linda Lotspeich, and Allan L. Reiss. White matter structure in autism: Preliminary evidence from diffusion tensor imaging. *Biological Psychiatry*, 55(3):323–326, 2004. ISSN 00063223. doi: 10.1016/j.biopsych.2003.10.022.

Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. On the syllabification of phonemes. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–316. Association for Computational Linguistics, 2009. ISBN 9781932432411.

Kamil Bartoń. *MuMIn: Multi-Model Inference*, 2016.

Douglas M. Bates, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, 2008. ISSN 01676393. doi: 10.1016/j.specom.2008.01.002.

Jennifer Urbano Blackford. Statistical Issues in Developmental Epidemiology and Developmental Disabilities Research: Confounding Variables, Small Sample Size, and Numerous Outcome Variables. *International Review of Research in Mental Retardation*, 33:93–120, 2006. ISSN 00747750. doi: 10.1016/S0074-7750(06)33005-4.

Jennifer Urbano Blackford. Propensity scores: method for matching on multiple variables in down syndrome research. *Intellectual and developmental disabilities*, 47(5):348–357, 2009. ISSN 1934-9491. doi: 10.1352/1934-9556-47.5.348.

Nathan Bodenstab and Aaron Dunlop. BUBS parser, 2011.

Paul Boersma and David Weenink. Praat: doing phonetics by computer, 2009.

Daniel Bone, Chi-chun Lee, Alexandros Potamianos, and Shrikanth Narayanan. An Investigation of Vocal Arousal Dynamics in Child-Psychologist Interactions using Synchrony Measures and a Conversation-based Model. In *Interspeech*, 2014a.

Daniel Bone, Chi-Chun Chun Lee, and Shrikanth Narayanan. Robust Unsupervised Arousal Rating: A rule-based framework with knowledge-inspired vocal features. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, 5:1–14, 2014b. ISSN 19493045. doi: 10.1109/TAFFC.2014.2326393.

Daniel Bone, Matthew P. Black, Anil Ramakrishna, Ruth B. Grossman, and Shrikanth S Narayanan. Acoustic-Prosodic Correlates of 'Awkward' Prosody in Story Retellings from Adolescents with Autism. In *Interspeech*, volume 2015-Janua, pages 1616–1620. International Speech and Communication Association, 2015.

Yoram S. Bonneh, Yoram Levanon, Omrit Dean-Pardo, Lan Lossos, and Yael Adini. Abnormal speech spectrum and increased pitch variability in young autistic children. *Frontiers in Human Neuroscience*, 4(January):1–7, jan 2011. ISSN 1662-5161. doi: 10.3389/fnhum.2010.00237.

George E. P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964. ISSN 0035-9246. doi: 10.2307/2287791.

MM Margaret M. Bradley and Peter J. PJ Lang. Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

John F. Brandt, Kenneth F. Ruder, and Jr. Shipp, Thomas. Vocal Loudness and Effort in Continuous Speech. *Journal of the Acoustical Society of America*, 46(6):1543–1548, 1969. ISSN 00014966. doi: 10.1121/1.1911899.

M. Brookes. Voicebox: Speech processing toolbox for Matlab, 2011.

Kate Bunton, Ray D. Kent, Jane F. Kent, and John C. Rosenbek. Perceptuo-acoustic assessment of prosodic impairment in dysarthria. *Clinical Linguistics & Phonetics*, 14(1):13–24, 2000. ISSN 0269-9206. doi: 10.1080/026992000298922.

Nick Campbell and Parham Mokhtari. Voice Quality: the 4th Prosodic Dimension. In *International Congress of Phonetic Sciences (ICPhS)*, pages 2417–2420, 2003. ISBN 1876346485.

Thomas F. Campbell and Christine Dollaghan. A method for obtaining listener judgments of spontaneously produced language: Social validation through direct magnitude estimation. *Topics in Language Disorders*, 12(2):42–55, 1992.

Pauline Chaste and Marion Leboyer. Autism risk factors: genes, environment, and gene-environment interactions. *Clinical research*, pages 281–292, 2012.

Nicholas Paul Chown. *A treatise on language methods and language-games in autism.* Phd dissertation, Sheffield Hallam University, 2012.

Joanne Cleland, Fiona E. Gibbon, Susan Peppé, Anne O'Hare, and Marion Rutherford. Phonetic and Phonological Errors in Children with High Functioning Autism and Asperger Syndrome. *International journal of Speech-Language Pathology*, 2010.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley-interscience, second edition, 2006.

David Crystal. Persevering with prosody. *International Journal of Speech-Language Pathology*, 11 (4):257, 2009. ISSN 1754-9507. doi: 10.1080/17549500902858753.

Adam W. Darkins, Victoria A. Fromkin, and D. Frank Benson. A characterization of the prosodic loss in Parkinson's disease. *Brain and Language*, 34(2):315–327, jul 1988. doi: 10.1016/0093-934X(88)90142-3.

Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002. ISSN 00014966. doi: 10.1121/1.1458024.

Céline De Looze and Stéphane Rauzy. Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2919–2922, 2009. ISSN 19909772.

Rajeev H. Dehejia and Sadek Wahba. Propensity Score Matching Methods for Non-Experimental Causal Studies. *Review of Economics and Statistics*, 84(November):151–161, 2002.

Julie Demouy, Monique Plaza, Jean Xavier, Fabien Ringeval, Mohamed Chetouani, Didier Périsse, Dominique Chauvin, Sylvie Viaux, Bernard Golse, David Cohen, and Laurence Robel. Differential language markers of pathology in Autism, Pervasive Developmental Disorder Not Otherwise Specified and Specific Language Impairment. *Research in Autism Spectrum Disorders*, 5(4): 1402–1412, 2011. ISSN 17509467. doi: 10.1016/j.rasd.2011.01.026.

Joshua John Diehl and Rhea Paul. The assessment and treatment of prosodic disorders and neurological theories of prosody. *International journal of Speech-Language Pathology*, 45(4): 287–292, 2009. ISSN 1754-9515. doi: 10.1080/17549500902971887.

Joshua John Diehl and Rhea Paul. Acoustic and perceptual measurements of prosody production on the profiling elements of prosodic systems in children by children with autism spectrum disorders. *Applied Psycholinguistics*, 34(01):1–27, 2011. ISSN 0142-7164. doi: 10.1017/S0142716411000646.

Joshua John Diehl and Rhea Paul. Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 6(1):123–134, 2012. ISSN 1750-9467. doi: 10.1016/j.rasd.2011.03.012.

Joshua John Diehl, Duane Watson, Loisa Bennetto, Joyce McDonough, and Christine Gunlogson. An acoustic analysis of prosody in high-functioning autism. *Applied Psycholinguistics*, 30(03): 385–404, 2009.

Ming Dong and Ravi Kothari. Look-ahead based fuzzy decision tree induction. *IEEE Transactions on fuzzy systems*, 9(3):461–468, 2001.

Lloyd M. Dunn and Douglas M. Dunn. *The British picture vocabulary scale*. GL Assessment Limited, 2009.

Keelan Evanini, Catherine Lai, and Klaus Zechner. The importance of optimal parameter setting for pitch extraction. *The Journal of the Acoustical Society of America*, 11(1):060004, 2010. ISSN 00014966. doi: 10.1121/1.3508047.

Thomas Ewender and Beat Pfister. Accurate Pitch Marking for Prosodic Modification of Speech Segments. *Interspeech*, pages 178–181, 2010.

Bruno Facon, David Magis, and John M. Belmont. Beyond matching on the mean in developmental disabilities research. *Research in Developmental Disabilities*, 32(6):2134–2147, 2011. ISSN 08914222. doi: 10.1016/j.ridd.2011.07.029.

Marisa Gomes Filipe. *Prosodic Abilities in Typically Developing Children and Those Diagnosed with Autism Spectrum Disorders*. Phd dissertation, Universidade do Porto Faculdade, 2014.

Marisa Gomes Filipe, Sónia Frota, São Luís Castro, and Selene G. Vicente. Atypical Prosody in Asperger Syndrome: Perceptual and Acoustic Measurements. *Journal of Autism and Developmental Disorders*, pages 1–10, mar 2014. ISSN 1573-3432. doi: 10.1007/s10803-014-2073-2.

J R Fontaine, Klaus R. Scherer, E B Roesch, and P Ellsworth. The World of Emotions is Not Wwo-Dimensional. *Psychological Science*, 18(12):1050–1057, 2007.

Hiroya Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and psychological interpretations. *STL-QPSR*, 22(1):1–20, 1981. ISSN 11045787. doi: citeulike-article-id:2912864.

Terisa P Gabrielsen, Megan Farley, Leslie Speer, and Michele Villalobos. Identifying Autism in a Brief Observation. *Pediatrics*, 135(2):e330–e338, 2015. doi: 10.1542/peds.2014-1428.

Michael L. Ganz. The Lifetime Distribution of the Incremental Societal Costs of Autism. *Archives of pediatrics & adolescent medicine*, 161(4):343–349, 2007. ISSN 1072-4710. doi: 10.1001/archpedi.161.4.343.

Krasimira Genova and Vassil Guliashki. Linear Integer Programming Methods and Approaches – A Survey. *Cybernetics and Information Technologies*, 11(1):3–25, 2011.

Daniel Gerlanc and Kris Kirby. *bootES: Bootstrap Effect Sizes*, 2015.

Mohammad Ghaziuddin and Leonore Gerstein. Pedantic speaking style differentiates asperger syndrome from high-functioning autism. *Journal of Autism and Developmental Disorders*, 26 (6):585–595, 1996. ISSN 0162-3257. doi: 10.1007/BF02172348.

Kyle B. Gorman, Jonathan Howell, and Michael Wagner. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193, 2011. ISSN 2291-1391.

Katherine Gotham, Susan Risi, Andrew Pickles, and Catherine Lord. The autism diagnostic observation schedule: Revised algorithms for improved diagnostic validity. *Journal of Autism and Developmental Disorders*, 37(4):613–627, 2007. ISSN 01623257. doi: 10.1007/s10803-006-0280-1.

Katherine Gotham, Andrew Pickles, and Catherine Lord. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 39 (5):693–705, 2009. ISSN 01623257. doi: 10.1007/s10803-008-0674-3.

Jonathan Green, Tony Charman, Andrew Pickles, Ming W. Wan, Mayada Elsabbagh, Vicky Slonims, Carol Taylor, Janet Mcnally, Rhonda Booth, Teodora Gliga, Emily J. H. Jones, Clare Harrop, Rachael Bedford, Mark H. Johnson, and the BASIS Team. Parent-mediated intervention versus no intervention for infants at high risk of autism: a parallel, single-blind, randomised trial. *The Lancet Psychiatry*, 0366(14):1–8, 2015. ISSN 2215-0366. doi: 10.1016/S2215-0366(14)00091-1.

Ruth B. Grossman, Lisa R Edelson, and Helen Tager-Flusberg. Emotional Facial and Vocal Expressions during Story Retelling by Children and Adolescents with High-Functioning Autism. *Journal of Speech, Language, and Hearing Research*, 56(3):1035–1044, 2013. ISSN 1558-9102. doi: 10.1044/1092-4388(2012/12-0067)Journal.

Xing Sam Gu and Paul R. Rosenbaum. Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993. ISSN 1061-8600. doi: 10.1080/10618600.1993.10474623.

Michele Gubian. Functional data analysis for speech research. *Online*, pages 1–4, 2013.

Michele Gubian, Francesco Cangemi, and Lou Boves. Automatic and Data Driven Pitch Contour Manipulation with Functional Data Analysis. *Analysis*, 2010.

Michele Gubian, Lou Boves, and Francesco Cangemi. Joint Analysis of F0 and Speech Rate with Functional Data Analysis. *International Conference of Acoustics, Speech and Signal Processing*, 2:4972–4975, 2011.

J. Hallmayer, S. Cleveland, a. Torres, J. Phillips, B. Cohen, T. Torigoe, J. Miller, a. Fedele, J. Collins, K. Smith, L. Lotspeich, Lisa A. Croen, S. Ozonoff, C. Lajonchere, J. K. Grether, and N. Risch. Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. *Archives of General Psychiatry*, 68(11):1095–1102, 2011. ISSN 0003-990X. doi: 10.1001/archgenpsychiatry.2011.76.

Ben B. Hansen. {Optmatch}: Flexible, Optimal Matching for Observational Studies. *R News*, 7 (2):18–24, 2007.

Francesca G. E. Happé. Editorial: Time to give up on Autism Spectrum Disorder? *Autism Research*, 10(January):10–14, 2017. doi: 10.1002/aur.1746.

Francesca G. E. Happé, Angelica Ronald, and Robert Plomin. Time to give up on a single explanation for autism. *Nature Neuroscience*, 9(10):1218–1220, 2006. doi: 10.1038/nn1770.

Isabelle Hesling, Bixente Dilharreguy, Susan Peppé, Marion Amirault, Manuel Bouvard, and Michèle Allard. The integration of prosodic speech in high functioning autism: a preliminary fMRI study. *PloS ONE*, 5(7):1–9, 2010.

Alison Presmanes Hill, Jan P. H. van Santen, Kyle B. Gorman, Beth Hoover Langhorst, and Eric Fombonne. Memory in language-impaired children with and without autism. *Journal of Neurodevelopmental Disorders*, 7(1):19, 2015a. ISSN 1866-1947. doi: 10.1186/s11689-015-9111-z.

Alison Presmanes Hill, Katharine E. Zuckerman, and Eric Fombonne. Epidemiology of Autism Spectrum Disorders. In Maria de los Angeles Robinson-Agramonte, editor, *Translational Approaches to Autism Apectrum Disorder*, pages 13–38. Springer International Publishing, 2015b. ISBN 9783319163215. doi: 10.1007/978-3-319-16321-5.

Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, 42(617):1–28, 2011. ISSN 1548-7660.

David C. Hoaglin. John W. Tukey and Data Analysis. *Statistical Science*, 18(3):311–318, 2003. ISSN 0883-4237. doi: 10.1214/ss/1076102418.

John M. Hoenig and Dennis M. Heisey. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, 55(1):19–24, 2001. ISSN 0003-1305. doi: 10.1198/000313001300339897.

Zoe Hopkins, Nicola Yuill, and Bill Keller. Children with autism align syntax in natural conversation. *Applied Psycholinguistics*, 37(02), 2016. ISSN 0142-7164. doi: 10.1017/S0142716414000599.

John-paul Hosom. F0 Estimation for Adult and Children's Speech. In *Eurospeech*, pages 317–320, 2005.

Pei-yun Hsueh, Prem Melville, and Vikas Sindhwani. Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. In *NAACL HLT Workshop on Active Learning for Natural Language Processing*, pages 27–35, 2009. doi: 10.1.1.157.5154.

Kathleen Hubbard and Doris A. Trauner. Intonation and emotion in autistic spectrum disorders. *Journal of Psycholinguistic Research*, 36(2):159–73, mar 2007. ISSN 0090-6905. doi: 10.1007/s10936-006-9037-4.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality Management on Amazon Mechanical Turk. In *ACM SIGKDD Workshop on Human Computation*, page 64, New York, USA, 2010. ACM Press. ISBN 9781450302227. doi: 10.1145/1837885.1837906.

Panagiotis G. Ipeirotis, Foster Provost, Victor S. Sheng, and Jing Wang. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28(2):402–441, mar 2013. ISSN 1384-5810. doi: 10.1007/s10618-013-0306-1.

Panos Ipeirotis. Crowdsourcing using Mechanical Turk: Quality Management and Scalability, 2011.

Mika Ito. *The contribution of voice quality to the expression of politeness: an experimental study.* PhD thesis, University of Edinburgh, 2005.

Christopher Jarrold and Jon Brock. To Match or Not to Match? Methodological Issues in Autism-Related Research. *Journal of Autism and Developmental Disorders*, 34(1):81–86, 2004. ISSN 01623257. doi: 10.1023/B:JADD.0000018078.82542.ab.

Anna Järvinen-Pasley, Susan Peppé, Gavin King-Smith, and Pamela Heaton. The relationship between form and function level receptive prosodic abilities in autism. *Journal of autism and developmental disorders*, 38(7):1328–40, aug 2008. ISSN 0162-3257. doi: 10.1007/s10803-007-0520-z.

Lisa N. Jefferies, Daniel Smilek, Eric Eich, and James T. Enns. Emotional Valence and Arousal Interact in Attentional Control. *Psychological science*, 19(3):290–5, 2008. ISSN 0956-7976. doi: 10.1111/j.1467-9280.2008.02082.x.

Keith Johnson. *Acoustic and Auditory Phonetics*. Blackwell, Oxford, second edition, 2003. doi: 10.1159/000078663.

Alexander Kain and Jan P. H. van Santen. Frequency-domain delexicalization using surrogate vowels. *Interspeech*, pages 474–477, 2010.

Constantijn Kaland, Marc Swerts, and Emiel Krahmer. Accounting for the listener: comparing the production of contrastive intonation in typically-developing speakers and speakers with autism. *The Journal of the Acoustical Society of America*, 134(3), 2013. ISSN 1520-8524. doi: 10.1121/1.4816544.

Leo Kanner. Autistic disturbances of affective contact. *Pathology*, pages 217–250, jan 1943. ISSN 00016586. doi: 10.1105/tpc.11.5.949.

Leo Kanner. Kugellmass, I. Newton: The Autistic Child. *Book Reviews*, pages 369–370, 1970.

Hideki Kawahara, Alain de Cheveigné, Hideki Banno, Toru Takahashi, and Toshio Irino. Nearly Defect-Free F0 Trajectory Extraction for Expressive Speech Modifications Based on STRAIGHT. In *Interspeech*, pages 537–540, 2005.

Tae-Hwan Kim and Halbert White. On More Robust Estimation of Skewness and Kurtosis: Simulation and Application to the S&P500 Index. *Finance Research Letters*, 1(1):56–73, 2004.

Géza Kiss and Jan P. H. van Santen. Estimating speaker-specific intonation patterns using the linear alignment model. In *Interspeech*. International Speech and Communication Association, 2013.

Géza Kiss, Jan P. H. van Santen, Emily Tucker Prud'hommeaux, and Lois M. Black. Quantitative Analysis of Pitch in Speech of Children with Neurodevelopmental Disorders. In *Interspeech*, pages 1343–1346, 2012. ISBN 9781622767595.

Géza Kiss, Kyle B. Gorman, and Jan P. H. van Santen. Selecting Statistically Similar Research Groups with ldamatch. *Under preparation*, 2017.

Esther Klabbers, A Kain, and Jan P. H. van Santen. Evaluation of speaker mimic technology for personalizing SGD voices. *Interspeech*, pages 1–4, 2010.

B Klaus and K Strimmer. fdrtool: Estimation of (local) false discovery rates and higher criticism. *CRAN*, 3(0), 2013.

Ami Klin, Celine A. Saulnier, Sara S. Sparrow, Domenic V. Cicchetti, Fred R. Volkmar, and Catherine Lord. Social and Communication Abilities and Disabilities in Higher Functioning Individuals with Autism Spectrum Disorders: The Vineland and the ADOS. *Journal of Autism and Developmental Disorders*, 37:748–759, 2007. doi: 10.1007/s10803-006-0229-4.

Ernst Kretschmer. *Physique and Character*. Springer, Berlin, 4th ed. edition, 1922.

Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. *lmerTest: Tests in Linear Mixed Effects Models*, 2016.

Øyvind Langsrud. Rotation tests. *Statistics and Computing*, 15(1):53–60, 2005. ISSN 09603174. doi: 10.1007/s11222-005-4789-5.

Øyvind Langsrud and Bjørn-Helge Mevik. *ffmanova: Fifty-fifty MANOVA*, 2012.

Marie-thérèse Le Normand, Sarah Boushaba, and Anne Lacheret-Dujour. Prosodic disturbances in autistic children speaking French. In *Speech Prosody*, pages 195–198, 2008.

Li Ching Lee, Rebecca A. Harrington, Brian B. Louie, and Craig J. Newschaffer. Children with autism: Quality of life and parental concerns. *Journal of Autism and Developmental Disorders*, 38(6):1147–1160, 2008. ISSN 01623257. doi: 10.1007/s10803-007-0491-0.

Rivka Levitan and Julia Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 3081–3084, 2011. ISSN 19909772. doi: 10.1.1.296.8560.

Rivka Levitan, Agustín Gravano, and Julia Hirschberg. Entrainment in Speech Preceding Backchannels. In *Annual Meeting of the Association for Computational Linguistics*, pages 113–117, 2011. ISBN 978-1-932432-88-6.

Ovsanna T Leyfer, Helen Tager-flusberg, Michael Dowd, J Bruce Tomblin, and Susan E. Folstein. Overlap Between Autism and Specific Language Impairment: Comparison of Autism Diagnostic Interview and Autism Diagnostic Observation Schedule Scores. *Autism Research*, 1(5):284–296, 2008. doi: 10.1002/aur.43.

Christopher H Lin, Mausam, and Daniel S Weld. Crowdsourcing Control: Moving Beyond Multiple Choice. *arXiv preprint arXiv:1210.4870*, 2012.

Chao Liu and Yi-Min Wang. TrueLabel + Confusions: A Spectrum of Probabilistic Models in Analyzing Multiple Ratings. In *International Conference on Machine Learning*, 2012.

Walter Loban. Language Development: Kindergarten through Grade Twelve. Technical report, NCTE Committee on Research, Urbana, Illinois, 1976.

Catherine Lord. Autism: From Research To Practice. *The American Psychologist*, 65(8):815–826, 2010. ISSN 1935-990X. doi: 10.1037/0003-066X.65.8.815.Autism.

Catherine Lord, S. Risi, and L. Lambrecht. The Autism Diagnostic Observation Schedule —Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 2000.

Catherine Lord, Susan Risi, Pamela S. DiLavore, Cory Shulman, Audrey Thurm, and Andrew Pickles. Autism from 2 to 9 years of age. *Archives of General Psychiatry*, 63(6):694–701, 2006. ISSN 0003-990X. doi: 10.1001/archpsyc.63.6.694.

T Loucas, T. Charman, Andrew Pickles, Emily Simonoff, Susie Chandler, David Meldrum, and Gillian Baird. Autistic symptomatology and language ability in autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry*, 49(11):1184–1192, 2008.

Heather MacFarlane, Kyle B. Gorman, Rosemary Ingham, Alison Presmanes Hill, Katina Papadakis, Géza Kiss, and Jan P. H. van Santen. Quantitative Analysis of Disfluency in Children with Autism Spectrum Disorders or Language Impairment. *PLoS ONE*, 12(3):e0173936, 2017.

Johnny L. Matson and Mary Shoemaker. Intellectual disability and its relationship to autism spectrum disorders. *Research in Developmental Disabilities*, 30(6):1107–1114, 2009. ISSN 08914222. doi: 10.1016/j.ridd.2009.06.003.

Susan Dickerson Mayes and Susan L. Calhoun. Impact of IQ, age, SES, gender, and race on autistic symptoms. *Research in Autism Spectrum Disorders*, 5(2):749–757, apr 2011. ISSN 17509467. doi: 10.1016/j.rasd.2010.09.002.

Susan Dickerson Mayes, Susan L. Calhoun, Michael J. Murray, Jill D. Morrow, Kirsten K. L. Yurich, Fauzia Mahr, Shiyoko Cothren, Heather Purichia, James N. Bouder, and Christopher Petersen. Comparison of scores on the Checklist for Autism Spectrum Disorder, Childhood Autism Rating Scale, and Gilliam Asperger's Disorder Scale for Children with Low Functioning Autism, High Functioning Autism, Asperger's Disorder, ADHD, and Typical Development.

Journal of Autism and Developmental Disorders, 39(12):1682–1693, 2009. ISSN 01623257. doi: 10.1007/s10803-009-0812-6.

Ashlynn McAlpine. *Prosodic Characteristics Observed in Verbal Children with Autism.* Master of science, Auburn University, 2012.

Ashlynn McAlpine, Laura W. Plexico, Allison M. Plumb, and Julie Cleary. Prosody in Young Verbal Children With Autism Spectrum Disorder. *Contemporary Issues in Communication Science and Disorders*, 41:120–132, 2014.

Joanne McCann and Susan Peppé. Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders*, 38(4):325–350, 2003. ISSN 1368-2822. doi: 10.1080/1368282031000154204.

Joanne McCann, Marion Rutherford, Susan Peppé, Fiona E. Gibbon, and Anne O'Hare. Prosody and Children with Autism. *Communication (The magazine of the National Autistic Society); NAS International Conference*, 39(1):43–35, 2005.

Joanne McCann, Susan Peppé, Fiona E. Gibbon, Anne O'Hare, and Marion Rutherford. Prosody and its relationship to language in school-aged children with high-functioning autism. *International Journal of Language & Communication Disorders*, 42(6):682–702, 2007. ISSN 1368-2822. doi: 10.1080/13682820601170102.

Joanne McCann, Susan Peppé, Fiona E. Gibbon, Anne O'Hare, and Marion Rutherford. The Prosody-Language Relationship in Children with High-Functioning Autism. In *Autism: An Integrated View from Neurocognitive, Clinical, and Intervention Research*, pages 214–235. Oxford: Blackwell Publishing, 2008.

Carol A. McGonegal, Lawrence R. Rabiner, and Aaron E. Rosenberg. A subjective evaluation of pitch detection methods using LPC synthesized speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):221–229, 1977. ISSN 0096-3518. doi: 10.1109/TASSP.1977.1162957.

Jane L. McSweeny and Lawrence D. Shriberg. Clinical research with the prosody-voice screening profile. *clinical linguistics & phonetics*, 15(7):505–528, 2001.

Carolyn B. Mervis and Bonita P. Klein-Tasman. Methodological Issues in Group-Matching Designs: alpha Levels for Control Variable Comparisons and Measurement Characteristics of Control and Target Variables. *Journal of Autism and Developmental Disorders*, 34(1):7–17, 2004. ISSN 01623257. doi: 10.1023/B:JADD.0000018069.69562.b8.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017.

Jon Miller and Robin Chapman. Systematic Analysis of Language Transcripts. *Madison, WI: Language Analysis Laboratory*, 1985.

Taniya Mishra. *Decomposition of fundamental frequency contours in the general superpositional intonation model Frequency Contours in the General*. Phd dissertation, Oregon Health and Science University, 2008.

Inger Moen. Deviant prosody in patients with cortical damage. *International Journal of Speech-Language Pathology*, 11(4):272–276, 2009. ISSN 1754-9507. doi: 10.1080/17549500902952507.

D Murphy and J Cutting. Prosodic comprehension and expression in schizophrenia. *Journal of neurology, neurosurgery, and psychiatry*, 53(9):727–30, 1990. ISSN 0022-3050. doi: 10.1136/jnnp.53.9.727.

Aparna Nadig and Holly Shaw. Acoustic and Perceptual Measurement of Expressive Prosody in High-functioning Autism: Increased Pitch Range and What It Means to Listeners. *Journal of Autism and Developmental Disorders*, 42(4):499–511, apr 2012. ISSN 1573-3432. doi: 10.1007/s10803-011-1264-3.

Aparna Nadig and Holly Shaw. Acoustic marking of prominence: how do preadolescent speakers with and without high-functioning autism mark contrast in an interactive task? *Language, Cognition and Neuroscience*, 30(1-2):32–47, dec 2015. ISSN 2327-3798. doi: 10.1080/01690965.2012.753150.

NIDCD National Institute on Deafness and Other Communication Disorders. Specific Language Impairment, 2016.

George L Nemhauser and Laurence A. Wolsey. Integer programming and combinatorial optimization. *Constraint Classification for Mixed Integer Programming Formulations*, 20:8–12, 1988.

K. Odell, Malcolm R. McNeil, John C. Rosenbek, and Linda Hunter. Perceptual characteristics of vowel and prosody production in apraxic, aphasic, and dysarthric speakers. *Journal of Speech and Hearing Research*, 34(1):67–80, 1991. ISSN 0022-4685.

Marguerite E. O'Haire, Samantha J. McKenzie, Alan M. Beck, and Virginia Slaughter. Animals May Act as Social Buffers: Skin Conductance Arousal in Children With Autism Spectrum

Disorder in a Social Context. *Developmental Psychobiology*, 57(5):584–595, 2015. ISSN 00121630. doi: 10.1002/dev.21310.

Jeanne M Paccia and Frank Curcio. Language processing and forms of immediate echolalia in autistic children. *Journal of Speech and Hearing Research*, 25(1):42–47, 1982. ISSN 0022-4685.

Barbara G. Parkhurst and Harry Levitt. The effect of selected prosodic errors on the intelligibility of deaf speech. *Journal of Communication Disorders*, 11(2-3):249–256, apr 1978. doi: 10.1016/0021-9924(78)90017-5.

Rhea Paul, Amy Augustyn, Ami Klin, and Fred R. Volkmar. Perception and Production of Prosody by Speakers with Autism Spectrum Disorders. *Journal of Autism And Developmental Disorders*, 35(2):205–220, apr 2005a. ISSN 0162-3257. doi: 10.1007/s10803-004-1999-1.

Rhea Paul, Lawrence D. Shriberg, Jane L McSweeny, Domenic V. Cicchetti, Ami Klin, and Fred R. Volkmar. Brief Report: Relations between Prosodic Performance and Communication and Socialization Ratings in High Functioning Speakers with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 35(6):861–869, 2005b. doi: 10.1007/s10803-005-0031-8.

Rhea Paul, Nancy Bianchi, Amy Augustyn, Ami Klin, and Fred R. Volkmar. Production of syllable stress in speakers with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 2 (1):110–124, 2008. ISSN 17509467. doi: 10.1016/j.rasd.2007.04.001.

Susan Peppé. Functionality and perceived atypicality of expressive prosody in children with Autism spectrum disorders. In *Speech Prosody*, pages 4–6, 2006.

Susan Peppé. Prosodic boundary in the speech of children with autism. In *International Congress of Phonetic Sciences (ICPhS)*, pages 1965–1968, 2007.

Susan Peppé. Why is prosody in speech-language pathology so difficult? *International Journal of Speech-Language Pathology*, 11(4):258–271, jan 2009. ISSN 1754-9507. doi: 10.1080/17549500902906339.

Susan Peppé and Joanne McCann. Assessing intonation and prosody in children with atypical language development: the PEPS-C test and the revised version. *Clinical Linguistics & Phonetics*, 17(4-5):345–354, jan 2003. ISSN 0269-9206. doi: 10.1080/0269920031000079994.

Susan Peppé, Jane Maxim, and Bill Wells. Prosodic variation in southern British English. *Language and speech*, 43(3):309–334, 2000. ISSN 0023-8309. doi: 10.1177/00238309000430030501.

Susan Peppé, J McCann, and Fiona E. Gibbon. Assessing prosodic and pragmatic ability in children with high-functioning autism. *Journal of Pragmatics*, 2006.

Susan Peppé, Joanne McCann, Fiona E. Gibbon, Anne O'Hare, and Marion Rutherford. Receptive and expressive prosodic ability in children with high-functioning autism. *Journal of Speech, Language, and Hearing Research*, 50(August 2007):1015–1028, 2007. ISSN 1092-4388. doi: 10.1044/1092-4388(2007/071).

Susan Peppé, Joanne Cleland, Fiona E. Gibbon, Anne O'Hare, and Pastora Martínez Castilla. Expressive prosody in children with autism spectrum conditions. *Journal of Neurolinguistics*, 24(1):41–53, 2011. ISSN 09116044. doi: 10.1016/j.jneuroling.2010.07.005.

Slav Petrov and Dan Klein. Learning and Inference for Hierarchically Split PCFGs. In *National conference on Artificial intelligence*, pages 1663–1666, 2007.

Wilbert Pronovost, M. Phillip Wakstein, D. Joyce Wakstein, and Albert T. Murphy. The Speech Behavior and Language Comprehension of Autistic Children. A Report of Research. Technical report, National Institution of Mental Health, Public Health Service, 1966.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

J O Ramsay, Hadley Wickham, Spencer Graves, and Giles Hooker. *fda: Functional Data Analysis*, 2014.

James O Ramsay, Giles Hooker, and Spencer Graves. *Functional data analysis with R and MATLAB*. Springer Science & Business Media, 2009.

C Radhakrishna Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948.

Okko Räsänen and Jouni Pohjalainen. Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. In *Interspeech*, pages 210–214, 2013.

Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2 (1):21–33, 2011. ISSN 9789673631575. doi: doi:10.1515/bile-2015-0008.

Brian Reichow, Peter Doehring, Domenic V. Cicchetti, and Fred R. Volkmar. Development, Procedures, and Application of the Evaluative Method for Determining Evidence-Based Practices in Autism. In Brian Reichow, Peter Doehring, Domenic V. Cicchetti, and Fred R. Volkmar, editors, *Evidence-Based Practices and Treatments for Children with Autism*, pages 1–408. Springer Science & Business Media, 2011. ISBN 9781441969736. doi: 10.1007/978-1-4419-6975-0.

Greg Ridgeway, Andrew R. Morral, Beth Ann Griffin, and Lane F. Burgette. Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG), 2014.

V. J. Roberts, S. M. Ingram, M. Lamar, and R. C. Green. Prosody impairment and associated affective and behavioral disturbances in Alzheimer's disease. *Neurology*, 47(6):1482–8, dec 1996.

Paul R. Rosenbaum and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.

Kenneth J Rothman, Sander Greenland, and Timothy L Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008.

Donald B. Rubin. Matching to Remove Bias in Observational Studies. *Biometrics*, 29(1):159–183, 1973.

Donald B. Rubin. Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. *Biometrics*, 47(4):1213–1234, 1991. ISSN 0006-341X. doi: 10.1017/CBO9780511810725.033.

Michael Rutter, Anthony Bailey, Catherine Lord, Carlo Cianchetti, and Giuseppina Sannio Fancello. *SCQ: Social Communication Questionnaire: Manuale*. Giunti OS, 2007.

Donald H. Saklofske, Lawrence G. Weiss, A. Lynne Beal, and Diane Coalson. The Wechsler Scales for assessing children's intelligence: past to present. In *Culture and children's intelligence: cross-cultural analysis of the WISC-III*. Academic Press, New York, 2003.

F. Santos, Nirit Brosh, Tiago H. Falk, Lonnie Zwaigenbaum, Susan E. Bryson, Wendy Roberts, Isabel M. Smith, Peter Szatmari, and Jessica A. Brian. Very Early Detection of Autism Spectrum Disorders Based on Acoustic Analysis of Pre-verbal Vocalizations of 18-month Old Toddlers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7567–7571, 2013. ISBN 9781479903566.

Klaus R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, 2005. ISSN 0539-0184. doi: 10.1177/0539018405058216.

B Schuller and L Devillers. Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm. *Interspeech*, pages 801–804, 2010.

Jasjeet S. Sekhon. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *Journal of Statistical Software*, 42(7):1–52, 2011.

Marsha Mailick Seltzer, Leonard Abbeduto, Marty Wyngaarden Krauss, Jan Greenberg, and April Swe. Comparison Groups in Autism Family Research: Down Syndrome, Fragile X Syndrome, and Schizophrenia. *Journal of Autism and Developmental Disorders*, 34(1):41–48, 2004. ISSN 01623257. doi: 10.1023/B:JADD.0000018073.92982.64.

Michal Shaked and Nurit Yirmiya. Matching Procedures in Autism Research: Evidence from Meta-Analytic Studies. *Journal of Autism and Developmental Disorders*, 34(1):35–40, 2004. ISSN 01623257. doi: 10.1023/B:JADD.0000018072.42845.83.

Megha Sharda, T. Padma Subhadra, Sanchita Sahay, Chetan Nagaraja, Latika Singh, Ramesh Mishra, Amit Sen, Nidhi Singhal, Donna Erickson, and Nandini C. Singh. Sounds of melody—Pitch patterns of speech in autism. *Neuroscience Letters*, 478(1):42–45, 2010. ISSN 03043940. doi: 10.1016/j.neulet.2010.04.066.

Shahla Sharifi, Zahra Azizi, and Mandana Nourbakhsh. The Tilt Model Acoustic Survey of Intonation in Children with Severe Autism. *International Journal of English Linguistics*, 6(4):78, 2016. ISSN 1923-8703. doi: 10.5539/ijel.v6n4p78.

Stephen J. Sheinkopf, Peter Mundy, D. Kimbrough Oller, and Michele Steffens. Vocal atypicalities of preverbal autistic children. *Journal of Autism and Developmental Disorders*, 30(4):345–354, 2000. ISSN 01623257. doi: 10.1023/A:1005531501155.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *KDD*, pages 614–622, 2008. ISBN 9781605581934.

Lawrence D. Shriberg and Carol J. Widder. Speech and prosody characteristics of adults with mental retardation. *Journal of Speech and Hearing Research*, 33(December):627–653, 1990.

Lawrence D. Shriberg, Joan Kwiatkowski, Carmen Rasmussen, Gregory L. Lof, and Jon F. Miller. The Prosody-Voice Screening Profile (PVSP): Psychometric data and reference information for children. Technical report, Waisman Center on Mental Retardation and Human Development, University of Wisconsin-Madison, 1992.

Lawrence D. Shriberg, Rhea Paul, Jane L. McSweeny, Ami Klin, Donald J. Cohen, and Fred R. Volkmar. Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. *Journal of Speech, Language, and Hearing Research*, 44(5): 1097–1115, oct 2001. ISSN 1092-4388. doi: 10.1044/1092-4388(2001/087).

Lawrence D. Shriberg, Rhea Paul, Lois M. Black, and Jan P. H. van Santen. The Hypothesis of Apraxia of Speech in Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 41:405–426, 2011. doi: 10.1007/s10803-010-1117-5.

Kåre Sjölander. The Snack Sound Toolkit, 2006.

Jeffrey A.. Smith and Petra E. Todd. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2 SPEC. ISS.):305–353, 2005. ISSN 03044076. doi: 10.1016/j.jeconom.2004.04.011.

David Snow and Heather L. Balog. Do children produce the melody before the words? A review of developmental intonation research. *Lingua*, 112(12):1025–1058, 2002. ISSN 00243841. doi: 10.1016/S0024-3841(02)00060-8.

Kemal Sönmez, Elizabeth Shriberg, Larry Heck, and Mitchel Weintraub. Modeling dynamic prosodic variation for speaker verification. In *ICSLP*, pages 3189–3192, 1998.

Grunya Efimovna Ssucharewa. Die schizoiden Psychopathien im Kindesalter. *Monatsschrift für Psychiatrie und Neurologie*, 60(3-4):235–261, 1926.

Vesna Stojanovik. Prosodic deficits in children with Down syndrome. *Journal of Neurolinguistics*, 24(2):145–155, 2011. ISSN 09116044. doi: 10.1016/j.jneuroling.2010.01.004.

Vesna Stojanovik and Jane Setter. Conditions in which prosodic impairments occur. *International journal of Speech-Language Pathology*, 11(4):293–297, 2009. ISSN 1754-9507. doi: 10.1080/17549500902943647.

Peter Szatmari, Lonnie Zwaigenbaum, and Susan E. Bryson. Conducting Genetic Epidemiology Studies of Autism Spectrum Disorders: Issues in Matching. *Journal of Autism and Developmental Disorders*, 34(1):49–57, 2004. ISSN 01623257. doi: 10.1023/B:JADD.0000018074.74369.cd.

Helen Tager-Flusberg. Strategies for Conducting Research on Language in Autism. *Journal of Autism and Developmental Disorders*, 34(1):75–80, 2004. ISSN 01623257. doi: 10.1023/B:JADD.0000018077.64617.5a.

David Talkin. A Robust Algorithm for Pitch Tracking (RAPT). In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pages 495–518. Elsevier Science, 1995.

Bruce Tomblin. Co-morbidity of autism and SLI: kinds, kin and complexity. *International Journal of Communication Disorders*, 46(2):127–137, 2011. doi: 10.1111/j.1460-6984.2011.00017.x.

Khiet P. Truong, David A. Van Leeuwen, and Franciska M. G. de Jong. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech Communication*, 54(9): 1049–1063, 2012. ISSN 01676393. doi: 10.1016/j.specom.2012.04.006.

Angela Tseng, Ravi Bansal, Jun Liu, Andrew J. Gerber, Suzanne Goh, Jonathan Posner, Tiziano Colibazzi, Molly Algermissen, I-Chin Chiang, James A. Russell, and Bradley S. Peterson. Using the Circumplex Model of Affect to Study Valence and Arousal Ratings of Emotional Faces by Children and Adults with Autism Spectrum Disorders. *Journal of autism and developmental disorders*, pages 1332–1346, nov 2013. ISSN 1573-3432. doi: 10.1007/s10803-013-1993-6.

Jan P. H. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8(2):95–128, 1994. ISSN 08852308. doi: 10.1006/csla.1994.1005.

Jan P. H. van Santen. NSF Report. Technical report, CSLU, 2014.

Jan P. H. van Santen and Bernd Möbius. A quantitative model of fo generation and alignment. In *Intonation*, pages 269–288. Springer, 2000.

Jan P. H. van Santen, Taniya Mishra, and Esther Klabbers. Estimating Phrase Curves in the General Superpositional Intonation Model. In *ISCA Workshop on Speech Synthesis*, pages 61–66, 2004.

Jan P. H. van Santen, Emily Tucker, Rhea Paul, Lois M. Black, and Lawrence Shriberg. Expressive Prosody in Autism: Effects of Prosody Function and Processing Demands. In *International Meeting for Autism Research*, 2008.

Jan P. H. van Santen, Emily Tucker Prud'hommeaux, and Lois M. Black. Automated Assessment of Prosody Production. *Speech Communication*, 51(11):1082–1097, 2009. ISSN 15378276. doi: 10.1016/j.specom.2009.04.007.Automated.

Jan P. H. van Santen, Emily Tucker Prud'hommeaux, Lois M. Black, and Margaret Mitchell. Computational prosodic markers for autism. *Autism*, 14(3):215–236, 2010. ISSN 8224406113. doi: 10.1093/biostatistics/manuscript-acf-v5.

Jan P. H. van Santen, Richard W. Sproat, and Alison Presmanes Hill. Quantifying repetitive speech in autism spectrum disorders and language impairment. *Autism Research*, 6(5):372–383, 2013. ISSN 19393792. doi: 10.1002/aur.1301.

Jing Wang, Panagiotis G. Ipeirotis, and Foster Provost. Quality-Based Pricing for Crowdsourced Workers. 2013.

Nigel Ward. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, 2004.

Simon K Warfield, Kelly H Zou, and William M Wells. Validation of Image Segmentation by Estimating Rater Bias and Variance. In R. Larsen, M. Nielsen, and J. Sporring, editors, *MICCAI 2006*, pages 839–847. Springer-Verlag Berlin Heidelberg, 2006. doi: 10.1007/11866763_103.

Lynn Waterhouse. ASD Validity. *Review Journal of Autism and Developmental Disorders*, 3(4): 302–329, 2016. doi: 10.1007/s40489-016-0085-x.

Lynn Waterhouse and Christopher Gillberg. Why autism must be taken apart. *Journal of Autism and Developmental Disorders*, 44(7):1788–1792, 2014. ISSN 15733432. doi: 10.1007/s10803-013-2030-5.

Sanford Weisberg. Yeo-Johnson Power Transformations. *Department of Applied Statistics, University of Minnesota*, 2001.

Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The Multidimensional Wisdom of Crowds. *Advances in Neural Information Processing Systems*, pages 1–9, 2010.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Advances in Neural Information Processing Systems*, 22:2035–2043, 2009.

Andrew J. O. Whitehouse, Matthew N. Cooper, Keely Bebbington, Gail Alvares, Ashleigh Lin, John Wray, and Emma J. Glasson. Evidence of a Reduction over Time in the Behavioral Severity of Autistic Disorder Diagnoses. *Autism Research*, 10(1):179–187, 2017. doi: 10.1002/aur.1740.

Hadley Wickham. Tidy Data. *Journal of Statistical Software*, 59(10):1–23, 2014.

Bodo Winter. Pseudoreplication in Phonetic Research. In *International Congress of Phonetic Sciences (ICPhS)*, pages 2137–2140, 2011.

Ericka L. Wodka, Pamela Mathy, and Luther Kalb. Predictors of Phrase and Fluent Speech in Children With Autism and Severe Language Delay. *Pediatrics*, 131(4):e1128–e1134, 2013. ISSN 1098-4275. doi: 10.1542/peds.2012-2221.

S Wolff. The first account of the syndrome Asperger described? *European Child & Adolescent Psychiatry*, 5(3):119–132, 1996. ISSN 1018-8827. doi: 10.1007/BF00571671.

Seyyed Pouria Fewzee Youssefi. *Affective Speech Recognition*. Phd dissertation, University of Waterloo, 2015.

B. Zei Pollerman. A place for prosody in a unified model of cognition and emotion. In *Speech Prosody*, 2002.