Automatic Vocal Recognition of a Child's Perceived Emotional State within the Speechome Corpus

by

Sophia Yuditskaya

S.B. Electrical Engineering and Computer Science (2002)
M.Eng. Electrical Engineering and Computer Science (2005)
Massachusetts Institute of Technology

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

Author_____
Sophia Yuditskaya
Program in Media Arts and Sciences
August 6, 2010

Certified by_____
Dr. Deb K. Roy
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by_____
Dr. Pattie Maes
Associate Professor of Media Technology
Associate Academic Head
Program in Media Arts and Sciences

Automatic Vocal Recognition of a Child's Perceived Emotional State within the Speechome Corpus

by

Sophia Yuditskaya

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on August 6, 2010 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Media Arts and Sciences

Abstract

With over 230,000 hours of audio/video recordings of a child growing up in the home setting from birth to the age of three, the Human Speechome Project has pioneered a comprehensive, ecologically valid observational dataset that introduces far-reaching new possibilities for the study of child development. By offering *In vivo* observation of a child's daily life experience at ultra-dense, longitudinal time scales, the Speechome corpus holds great potential for discovering developmental insights that have thus far eluded observation. The work of this thesis aspires to enable the use of the Speechome corpus for empirical study of emotional factors in early child development. To fully harness the benefits of Speechome for this purpose, an automated mechanism must be created to perceive the child's emotional state within this medium.

Due to the latent nature of emotion, we sought objective, directly measurable correlates of the child's perceived emotional state within the Speechome corpus, focusing exclusively on acoustic features of the child's vocalizations and surrounding caretaker speech. Using Partial Least Squares regression, we applied these features to build a model that simulates human perceptual heuristics for determining a child's emotional state. We evaluated the perceptual accuracy of models built across child-only, adult-only, and combined feature sets within the overall sampled dataset, as well as controlling for social situations, vocalization behaviors (e.g. crying, laughing, babble), individual caretakers, and developmental age between 9 and 24 months. Child and combined models consistently demonstrated high perceptual accuracy, with overall adjusted R-squared values of 0.54 and 0.58, respectively, and an average of 0.59 and 0.67 per month. Comparative analysis across longitudinal and socio-behavioral contexts yielded several notable developmental and dyadic insights. In the process, we have developed a data mining and analysis methodology for modeling perceived child emotion and quantifying caretaker intersubjectivity that we hope to extend to future datasets across multiple children, as new deployments of the Speechome recording technology are established. Such large-scale comparative studies promise an unprecedented view into the nature of emotional processes in early childhood and potentially enlightening discoveries about autism and other developmental disorders.

Thesis Supervisor: Deb K. Roy
Title: Associate Professor, Program in Media Arts and Sciences

Automatic Vocal Recognition of a Child's Perceived Emotional State within the Speechome Corpus

by

Sophia Yuditskaya

The following person served as a reader for this thesis:

Thesis Reader:_____

Dr. Rosalind W. Picard
Professor of Media Arts and Sciences
Director of Affective Computing Research
Co-Director, Autism Communication Technology Initiative
Program in Media Arts and Sciences

Automatic Vocal Recognition of a Child's Perceived Emotional State within the Speechome Corpus

by

Sophia Yuditskaya

The following person served as a reader for this thesis:

Thesis Reader:_____

Dr. Letitia R. Naigles
Professor of Psychology
Director, UConn Child Language Lab
University of Connecticut

Automatic Vocal Recognition of a Child's Perceived Emotional State within the Speechome Corpus

by

Sophia Yuditskaya

The following person served as a reader for this thesis:

Thesis Reader:_____

Dr. Matthew S. Goodwin
Research Scientist, Dept. of Psychiatry and Human Behavior, Brown University
Director of Clinical Research, Postdoctoral Fellow
Co-Director, Autism Communication Technology Initiative
MIT Media Laboratory

I would like to dedicate this thesis to the memory of Professor Edward G. Carr.

Acknowledgements

I would like to thank my advisor Professor Deb Roy for giving me the opportunity to pursue graduate studies at the MIT Media Laboratory and to work with the Speechome dataset. It is inspiring to think of the breakthroughs in developmental research that await discovery within this medium, and I am grateful to have had this opportunity to contribute my efforts and abilities to this vision.

I would like to express my heartfelt gratitude to Professor Rosalind Picard and Dr. Matthew Goodwin for their dedicated guidance and support, their unwavering confidence in me, and for keeping my best interest at heart. They have taught me so much about machine learning and statistical analysis for behavioral research, affective computing and the issues involved in any research involving emotion, and the significance of technology in the study and treatment of autism and other developmental disorders, all of which have come together to form the foundation of my thesis research. I am deeply grateful for all the feedback, advice, and opportunities for in-depth discussion that they both have provided in the course of my thesis work. I am also grateful to Professor Letitia Naigles for her critical feedback and valuable insights from the field of developmental psychology, and to my advisor Professor Deb Roy for guiding me to think about my work in terms of psychoacoustics.

Thanks also to Professor Pattie Maes, Professor Alex (Sandy) Pentland, Chris Schmandt, Ian Eslick, Sajid Sadi, and Ehsan Hoque for their critiques during the thesis proposal process. I am also grateful to Professor Mitchel Resnick for his feedback and support. I would also like to thank Benjamin Waber and Alex Kaufman for their guidance about correlation and regression, and for being sounding boards for the different statistical analysis strategies I had considered during my thesis work.

I am indebted to my seven wonderful UROP students, whose annotations made this thesis possible: Jennifer Bustamante, Christine M. Chen, Chloe Margarita Dames, Nicole Hoi-Cheng Fong, Jason Hoch, Emily R. Su, and Nahom Workie. You were dependable, conscientious, and inquisitive. With your help, the most challenging part of this thesis – data collection – went more smoothly than I ever imagined possible. I feel so lucky to have had all of you as my assistants.

I would like to thank my colleagues in the Cognitive Machines group: Philip DeCamp, Rony Kubat, Matt Miller, Brandon Roy, Stefanie Tellex, and Leo Tsourides for the prior Speechome work that has made my thesis possible; Jeff Orkin, Aithne Sheng-Ying Pao, Brandon Roy, Hilke Reckman, George Shaw, Leo Tsourides, and Soroush Vosoughi for their feedback and encouragement; and Tomite Kaname for being such a gracious and pleasant office-mate.

I am especially grateful to Brandon Roy, Matt Miller, Philip DeCamp, Soroush Vosoughi, and Hilke Reckman for their direct assistance in my thesis work. Brandon, thank you for always being willing to help with my questions regarding the Speechome database despite your heavy workload. Several times you dropped what you were doing to lend support in some

of my most critical times of need, such as when the Speechome network refused to work after we moved to E14, and when my annotation database disk crashed. Matt, thanks for sharing your speaker ID code, labels, and performance statistics, and for being so accessible in answering my questions as I applied them in my thesis. Philip, thank you for your assistance with memory and performance issues that I encountered in integrating your video playback libraries into my annotation interface. Soroush, thank you for acquainting me with the basics of Praat scripting. Hilke, thank you for offering your linguistics expertise to provide helpful tips and references in discussing ideas for thesis directions. It was really nice to know that you kept me in mind as you encountered and recalled relevant sources.

Outside of my group, I would like to also thank Angela Chang, Karen Brennan, Santiago Alfaro, Andrea Colaco, Micah Eckhardt, Elliott Hedman, Elly Jessop, Inna Koyrakh, and many others, for their cheerful congeniality, feedback, and encouragement.

I would also like to thank Jane Wojcik, Will Glesnes, Peter Pflanz, and Jon Ferguson at Necsys, who spent many hours in the machine room with (and without) Brandon and me, debugging the Speechome network during that crazy time in January 2010, and for helping me with numerous tech support issues throughout my past two years at the Media Lab. Thanks also to Joe Wood for his help as the new Cognitive Machines tech specialist, keeping the Speechome network and backups running smoothly since he came aboard, and for his efforts to salvage annotation data from my failed disk.

I would like to express my appreciation to Linda Peterson, Aaron Solle, Alex Khitrik and Karina Lundahl, Media Lab administrators whose various efforts in offering support, answering administrative questions, arranging meetings, and delivering thesis drafts and paperwork have made this thesis possible.

Throughout my graduate studies, dancesport has continued to be an important part of my life, keeping me sane and balanced through its unique combination of musical expression, collaborative partnership, and physical activity. To all my friends on the MIT Ballroom Dance Team, thank you for your companionship and encouragement, for helping me stay positive, for sharing in our collective love for this artform, and for being my role models on achieving both work and dance with discipline and rigor. Special thanks to my dance partner Will Phan for his patience and understanding, and for keeping my spirits up with his witty humor. To Armin Kappacher, who is so much more than a (brilliant) dance teacher, thank you for your support and warm words of encouragement. Your insightful wisdom gave me hope and helped me grow as a person in facing the adversities that surrounded my thesis work. Your selfless passion for teaching has inspired me to approach my work with the same kind of dedication and commitment to quality.

Finally, I wish to thank my family: this thesis stands as a testament to their unconditional love, constant encouragement, and moral support. Throughout my life, my mother Inessa Yuditskaya and my twin-sister Susan have taught me by example, through spiritual guidance, and with unwavering confidence in my abilities, to be a woman of strength, independence, and creativity in everything that I do. Knowing that they are always there for me and that I am loved so deeply is my source of joy, hope, and inner peace.

Table of Contents

Table of Figures

List of Tables

Chapter 1

Introduction

Originally developed to study child language acquisition, the Human Speechome Project (D. Roy, 2009; D. Roy et al., 2006) has pioneered a new kind of dataset – dense, longitudinal, ecologically-valid – that introduces far-reaching new possibilities for the study of child development. Consisting of over 230,000 hours of multichannel raw audio/video recordings, the Speechome corpus forms a comprehensive observational archive of a single typically developing child growing up in the home setting, starting from birth to the age of 3. Due to its scale, harnessing the benefits of this data requires the development of novel data mining strategies, manual data collection tools, analysis methods, and machine learning algorithms that can transform the raw recorded media into metadata, and ultimately insights, that are meaningful for answering developmental research questions. Metadata serving this purpose includes transcribed speech, speaker identity, locomotive trajectories, and head orientation, all of which represent ongoing efforts surrounding the Speechome corpus. The methods and technologies designed and implemented to date are notable advances, but collectively they only scratch the surface of what the Speechome corpus has to offer. Of note, there are currently no existing resources in the Speechome corpus for the study of affective research questions – those pertaining to the understanding of emotional factors and processes in development.

Understanding the nature of emotional processes in early childhood and the inclusion of emotion-related variables in studying other aspects of development are important themes in the study of child development. Supported by considerable theoretical and empirical work that describes emotion and cognition as "inseparable components of the developmental process" (Bell & Wolfe, 2004; Calkins & Bell, 2009), there is increasing evidence to suggest that emotions play a regulatory role in perception, learning, retrieval of memories, and organization of cognitive states (Dawson, 1991; K. W. Fischer et al., 1990; Trevarthen, 1993; Wolfe & Bell, 2007). Emotions are often described in developmental

psychology as organizers that shape behavior (Cole et al., 1994; K. W. Fischer et al., 1990; Trevarthen, 1993). "[Emotions] are a part of the dynamic generation of conscious, intelligent action that precedes, attracts, and changes experiences," Trevarthen writes (1993). Through conditioning, habitual patterns of emotion influence long-term development. Such emotional patterns comprise and define a child's temperament (Kagan, 1994; Wolfe & Bell, 2007), which has been found to correlate with outcomes in social competence (Lemerise & Arsenio, 2000), cognitive development, and language performance (Wolfe & Bell 2001). Temperament has also been linked to psychopathologies such as depression, bipolar disorder, drug abuse, and psychosis (Camacho & Akiskal, 2005; Rothbart, 2005; Sanson et al., 2009).

In attachment theory (Bowlby, 1973), the earliest bonds formed by infants with their mothers are thought to be central in influencing development. Here, emotion is said to organize the "security or insecurity of the mother-infant relationship, which is then internalized as a working model and carried into subsequent relations" (Cassidy, 1994; Cole et al., 1994; Simpson et al., 2007). Many works in infancy research confirm this theory by demonstrating that emotion organizes the development of social relations, physical experience, and attention (Cole et al., 1994; Klinnert et al., 1983; Rothbart & Posner, 1985; Sroufe et al., 1984).

Emotional processes are also believed to be critical in the development of language (Bloom, 1998; Bolnick et al., 2006; Ochs & Schieffelin, 1989; Trevarthen, 1993; Wolfe & Bell, 2007). On the one hand, emotions serve as motivators, as children learn language initially because they strive to express their internal experiences (Bloom, 1998; Bolnick et al., 2006; Ochs & Schieffelin, 1989). Social referencing, the act of monitoring the affective state of those who are speaking to us, is a skill that infants acquire early and use to discover meaning in what is said to them by caretakers (Ochs & Schieffelin, 1989). Ultimately, social referencing teaches the infant not only how to use language to express emotional state, but also how to understand the emotional state of others. On the other hand, there is evidence to suggest that emotion and language compete for cognitive energy (Bloom, 1998): children who spend more time in an affectively neutral state have been observed to show better language development. This suggests that the regulation of emotion, particularly the

child's developing ability to self-regulate, is also an important process in facilitating language acquisition.

Despite its importance, empirical progress in studying emotion during early childhood development has been elusive. Just as with language acquisition (D. Roy, 2009; D. Roy et al., 2006), empirical work in these areas has been constrained by the biases and limitations inherent in traditional experimental designs (Trevarthen, 1993), in which observational data is collected at a laboratory or by researchers visiting a child's home. Further, such designs naturally involve sparse longitudinal samples spaced weeks or months apart, adding up to mere snapshots of a child's development that offer "little understanding of the process of development itself " (Adolph et al., 2008; D. Roy, 2009). Such sparse sampling can misrepresent the actual course of development, as studies of children's physical growth have shown (Johnson et al., 1996; Lampl et al., 2001). While sampling every three months produces a smooth, continuous growth curve, sampling infants' growth at daily and weekly intervals reveals a non-continuous developmental process in which long periods of no growth are punctuated by short spurts (Adolph et al., 2008; Johnson et al., 1996; Lampl et al., 2001). Sparse sampling is also likely to produce large errors in estimating onset ages, which are useful in screening for developmental delay. By offering *In vivo* observation of a child's daily life experience at ultra-dense, longitudinal time scales even to the order of milliseconds, the Speechome corpus holds great potential for discovering developmental insights that have thus far eluded observation.

The work of this thesis is dedicated to enabling the use of Speechome's dense, longitudinal, ecologically valid observational recordings for the study of emotional factors in early child development. In any research question surrounding early emotional processes, the child's emotional state is a key variable for empirical observation, measurement, and analysis. Due to the immense scale of the Speechome corpus, these operations must be automated in order to fully harness the benefits of ultra-dense observation. Our main challenge is therefore to furnish an automated mechanism for determining the child's emotional state within the medium of Speechome's audio-visual recordings.

Like pitch and color, however, emotional state is not an objective physical property, but rather a subjective, highly perceptual construct. While there is an objective

physiological component that accompanies the "feeling" of an emotion, the experience itself is the perceptual interpretation of this feeling. For example, a physiological increase in heart rate and blood pressure often accompanies the emotional experience of "fear", as part of the reflexive "fight-or-flight" response of the autonomic nervous system to perceived threats or dangers (Goodwin et al., 2006). The outward, physical expression of this emotional experience becomes the medium through which others perceive a person's emotional state. Facial expressions, physiological measures such as heart rate and respiration, loudness and intonation of the voice, as well as bodily gestures, all provide objective physical clues that guide the observer in perceiving a person's latent internal emotional experience. The latent nature of emotion means that we can only speak of a person's *perceived* emotional state in the context of empirical observation.

In the Speechome recordings, the vocal and visual aspects of the child's outward emotional expressions are the main objective perceptual cues to the child's internal emotional state that are available for empirical study. In addition, caretaker response can also provide contextual clues, such as a mother providing comfort when the child is upset. To automate the observation, measurement, and analysis of the child's perceived emotional state in the Speechome corpus, we seek directly measurable correlates of the child's perceived emotional state among these objective perceptual cues. In this thesis, we focus exclusively on the acoustic attributes of the child's vocal emotional expression and surrounding adult speech. We apply these vocal correlates to build an emotion recognition model that simulates human perceptual heuristics for determining a child's emotional state. Creating separate models for specific longitudinal, dyadic, and situational subsets of the data, we also explore how correlations between perceived child emotion and vocal expression vary *longitudinally* during the 9 to 24 month developmental period, *dyadically* in relation to specific caretakers, and *situationally* based on several socio-behavioral contexts. In the process, we develop a data mining and analysis methodology for modeling perceived child emotion and quantifying intersubjectivity that we hope to extend to future datasets across multiple children, as new deployments of the Speechome recording technology are established.

One area of developmental research that has served as a specific inspiration for this work is autism. Emotional state is a construct with much relevance to the study and

treatment of autism. Chronic stress and arousal modulation problems are major symptomatic categories in the autistic phenotype, in addition to abnormal socio-emotional response (De Giacomo & Fombonne, 1998) that has been attributed to profound emotion dysregulation (Cole et al., 1994). Stress is associated with aversive responses; in particular, aversion to novel stimuli (Morgan, 2006). For this reason, it has been proposed that chronic stress and arousal modulation problems negatively affect an autistic infant's ability to engage in social interaction during early development (Morgan, 2006; Volkmar et al., 2004) and onwards. Missing out on these early interactions can lead to problems in emotional expressiveness, as well as in understanding and perceiving emotional expressions in others (Dawson, 1991).

Understanding the early processes by which such deficiencies emerge in autistic children can be a significant breakthrough towards early detection and developing effective therapies. At the same time, such insights from autism can also inform our understanding of neurotypical socio-emotional development:

> Because children with autism have deviant development, rather than simply delayed development, studies of their patterns of abilities offer a different kind of opportunity for disentangling the organization and sequence of [normal socio-emotional] development.
>
> (Dawson, 1991)

Forthcoming new deployments of the Speechome recording technology are part of a greater vision to launch large-scale comparative studies between neurotypical and autistic children using Speechome's dense, longitudinal, ecologically-valid datasets. Due to the salience of emotional factors in autism research, such comparative studies would seek to ask many affective research questions of the Speechome corpora. By modeling child emotion in the Speechome corpus for the purpose of creating an automated emotion recognition mechanism within this medium, we hope to prepare Speechome for the service of such questions, setting the stage for an unprecedented view into the nature of emotional processes in early childhood and potentially enlightening new discoveries about autism and other developmental disorders.

1.1 Related Work

With applications in human-computer interaction, developmental research, and early diagnosis of developmental disorders, there is a sizeable and growing body of work related to automatic recognition of emotional states via acoustic analysis of vocal attributes (Douglas-Cowie et al., 2003; Scherer, 2003; Ververidis & Kotropoulos, 2006). In our discussion of related work, we begin with the most general area of prior work in this field – emotion recognition from adult speech – and then proceed to works that are addressing emotion recognition in child vocalizations and related developmental research questions.

1.1.1 Emotion Recognition from Adult Speech

A large proportion of prior research in this area focuses on emotion recognition in adult speech, for the purpose of developing intelligent interfaces that can accurately understand the emotional state of the human speaker and act in accordance with this understanding. For example, Hansen and Cairns (1995) trained a speech recognition system for aircraft cockpits using stressed speech. Several ticket reservation systems have benefited from emotion recognition models that detect annoyance or frustration, changing their response accordingly (Ang et al., 2002; Schiel et al., 2002). Similar successes have been achieved in call center voice-control applications (Lee & Narayanan, 2005; Petrushin, 1999). France et al (France et al., 2000) have successfully applied emotion recognition of adult speech as a diagnostic tool in medical applications. In addition to all of the above examples, Ververidis and Kotropoulos (2006) note the relevance of emotional speech recognition methods to psychology research, in coping with the "bulk of enormous speech data" to systematically extract "speech characteristics that convey emotion" (Mozziconacci & Hermes, 2000; Ververidis & Kotropoulos, 2006).

Social robotics has also received attention in emotion recognition research (Batliner et al., 2006; Breazeal, 2001; Breazeal & Aryananda, 2002; Oudeyer, 2003). As we discuss further in Section 1.1.2, studies have also sought to develop emotion recognition systems using child speech, aimed at child-robot interaction (Batliner et al., 2006), with applications in the study and treatment of autism (Dautenhahn, 1999; Scassellati, 2005; Tapus et al.,

2007), among other things. The task of creating a social robot that can respond appropriately and flexibly to the affective state of its human companions in the course of natural conversation is particularly challenging, because the scope of emotions is unconstrained. In contrast, specialized aircraft cockpits and ticket reservation systems listed above only need to detect a small subset, such as stress, annoyance, and frustration. Applications in social robotics, and intelligent interfaces in general, require that an emotion recognition model generalize across speakers, since a single robot or interface may interact with many different people. In our case, however, we set out to create a model that is only relevant within the speech corpus of a single child. Each corpus would have its own model built separately using that corpus.

Because of the challenges involved in obtaining datasets of natural, spontaneous emotional speech recordings and applying them for focused experimental analysis, much of this work has called upon actors to simulate various specific emotions as needed by the goals of each study (Banse & Scherer, 1996; Burkhardt et al., 2005; Dellaert et al., 1996; Douglas-Cowie et al., 2003; Hozjan et al., 2002; Lay New et al., 2003; Lee et al., 2004; Navas et al., 2004; Seppanen et al., 2003). Each study differs based on the set of acoustic features they use, and the categories of emotion they classify. For high quality recordings enabling acoustic analysis, they record in a sound-proof studio, which intrinsically reduces spontaneity of recorded speech. Further, little is known about the relationship between acted and spontaneous everyday emotional speech (Douglas-Cowie et al., 2003), so the utility of results obtained using acted speech is questionable. At the very least, using this methodology misses out on the rich variability of subtle grades of emotion that occur during the course of natural conversation in real-life social situations. In spontaneous speech, canonical emotions such as happiness and anger occur relatively infrequently, and the distribution of emotion classes is highly unbalanced (Neiberg et al., 2006).

The inadequacy of acted-speech methodologies has led to many attempts at collecting and annotating naturalistic speech for the study of vocal emotion. Strategies have included:

- Fitting volunteers with long-term recorders that sample their daily vocal interactions (Campbell, 2001)

- Recording telephone conversations (Campbell, 2001) and call center dialogs (K. Fischer, 1999; Lee & Narayanan, 2005; Neiberg et al., 2006; Vidrascu & Devillers, 2005)

- Placing a microphone between subjects talking in an informal environment, such as at the dinner table to record conversations during family meals (Campbell, 2001), a desk in the workplace to record job interviews (Rahurkar & Hansen, 2002) and interactions between coworkers (Campbell, 2001), and in a doctor's office (France et al., 2000)

- Recording user sessions at the interface system for which the emotion recognition mechanism was being developed (Steiniger et al., 2002)

Other works have used movie sequences as well as radio and television clips as an improvement in naturalism over acted laboratory readings (Douglas-Cowie et al., 2000; Greasley et al., 1995; Roach et al., 1998; Scherer, 2003; Sharma et al., 2002). Recording quality, background noise, crosstalk, and overlapping speech are among the challenges that these works have encountered and tried to address (Douglas-Cowie et al., 2003; Truong & van Leeuwen, 2007).

A major bottleneck in emotion recognition research that uses naturalistic speech databases is the tedious, error-prone, and time-consuming manual annotation that is necessary to pick out and sort the salient exemplars according to the categories being studied, such as utterances corresponding to a particular emotion. In some cases, the subjects are asked, immediately after a recording session, to annotate the emotions that they felt during the recording process (Busso & Narayanan, 2008; Truong et al., 2008). In most other cases, annotation is done independently of the data collection process, by researchers.

Related to the annotation problem is the question of which emotion coding labels to include in the annotation taxonomy. Studies such as (Greasley et al., 1995; Greasley et al., 2000; Greasley et al., 1996) develop and evaluate coding schemes for annotating emotion in natural speech databases. Among the design parameters of a coding scheme, these studies examine parameters such as:

- Free-choice codings vs. fixed-choice codings (Greasley et al., 2000). In free-choice codings, annotators use their own words to describe the emotion of an utterance as they see fit. In fixed-choice codings, the annotator has to choose from a hard-coded set of labels.

- Categorical vs. dimensional categorization (Douglas-Cowie et al., 2003; Scherer, 2003). A categorical, or discrete, design subdivides the range of human emotions into a set of fundamental qualitative categories, such as *joy, contentment, interest, surprise, unease, anger,* and *pain* (Douglas-Cowie et al., 2003)**.** A dimensional design maps emotional states to numerical coordinates within a two- or three-dimensional space, such as those proposed by Schlosberg (1952), Osgood et al (1957), and Ortony et al (1988), and applied by many others (J. A. Bachorowski & Owren, 1995; Feldman-Barrett, 1995; Greasley et al., 2000). Each axis defining these spaces corresponds to an orthogonal attribute of emotional expression, such as *valence* – the pleasantness or unpleasantness of the emotion, and *arousal* – the degree of emphasis, power, or excitement in the emotion. Emotional utterances are then rated according to a graded scale along each of these axes.

- Granularity of categorized emotion (Douglas-Cowie et al., 2000; Stibbard, 2000) – the finer-grained the characterization, the more categories there are, and the less occurrences of each category in the dataset. However, fewer but more inclusive categories may aggregate multiple finer emotional states each having distinct acoustic properties, resulting in less successful recognition mechanisms when modeling them as a single super-emotion.

1.1.2 Emotion Recognition from Child Vocalizations

While adult speech has laid the groundwork of the field, more relevant to the work of this thesis are the many studies that have sought to develop emotion recognition systems for child vocalizations. We reference the following relevant themes that have influenced system designs and methodologies among this body of work:

- Detecting vocalizations that correspond to a specific single emotion, such as crying or laughing (Gustafson & Green, 1989; Nwokah et al., 1993; Ruvolo & Movellan, 2008). We discuss (Ruvolo & Movellan, 2008) in greater detail below (Section 1.1.2.1).

- Diagnosing developmental impairments based on the acoustic properties of a specific type of emotional vocalization, such as crying or laughing. (Fuller, 1991; Garcia & Garcia, 2003; Hudenko et al., 2009; Petroni et al., 1995; Petroni, Malowany, Johnston, & Stevents, 1994; Schonweiler et al., 1996; Varallyay Jr et al., 2007). We discuss (Garcia & Garcia, 2003; Hudenko et al., 2009; Petroni et al., 1995; Varallyay Jr et al., 2007) in greater detail below (Section 1.1.2.2).

- Distinguishing between emotional and communicative vocalizations (Papaeliou et al., 2002). We include (Papaeliou et al., 2002) in the discussion of Section 1.1.2.2.

- Distinguishing between two or more different emotional states (Batliner et al., 2006; Petrovich-Bartell et al., 1982; Scheiner et al., 2002; Shimura & Imaizumi, 1994). We discuss (Scheiner et al., 2002) and (Batliner et al., 2006) further in Section 1.1.2.3.

1.1.2.1 Cry Detection

Ruvolo and Movellan (2008) implemented robust cry detection to support the immersion of social robots in childhood education settings. A robust cry detection mechanism gives these robots the useful function of assisting teachers with managing classroom mood, by offering emotional support and stimulation to a crying child or alerting the teacher in more serious cases. The challenge of building a cry detector for this purpose is the noisy and unpredictable nature of the preschool setting – a natural attribute of naturalistic settings that is also shared by the Speechome corpus.

A full day of audio was recorded in the preschool environment, totaling 6 hours. Human annotators labeled 40 minutes of this audio corpus, 2-sec audio clips at a time,

according to whether a cry was present or not present in each clip. Spatio-Temporal Box filters were applied to spectrograms of each clip, forming the feature set for training and classification, with Gentle-Boost as the classification algorithm. The resulting cry detector achieved a classification accuracy of 94.7% for 2-sec audio clips and also found 8-sec clips to achieve even better accuracy of 97%.

1.1.2.2 Diagnosis and Study of Developmental Impairments

Hudenko et al (2009) found vocal acoustics of laughter to have discriminatory value in distinguishing between neurotypical and autistic children. This study was done on older children, ranging in age from three to ten years of age. Fifteen autistic children aged eight to ten were each matched, in part of the study, with neurotypical children sharing the same Verbal Mental Age, resulting in the large range in ages overall. Each child was recorded in individual laboratory sessions in which a Laugh-Assessment Sequence (LAS) was used to elicit laughter through playful interaction between examiner and child. Annotation of these recordings coded laugh utterances according to criteria established by (J. Bachorowski et al., 2001), with Cohen's kappa of 0.95. Various acoustic measures, such as voicing, laugh duration, number of "syllables", and pitch-related metrics were computed for each laugh utterance, and then evaluated using ANOVA for significance in distinguishing between neurotypical and autistic groups. Results of this analysis showed voicing to be a highly significant discriminatory feature: while neurotypical children exhibited both voiced and unvoiced laughter in significant proportions, autistic children primarily exhibited voiced laughter, with only a negligible amount of unvoiced laughter.

Varallyay Jr et al (2007) built on the work of Schonweiler (1996) and Wermke (2002) in characterizing developmental trends in the melody of infant cries to detect hearing impairments and central nervous system disorders in infants. A database of 2460 crying samples from 320 infants was collected as part of a 5-year-long data collection in several hospitals and homes. Although these were ecologically valid settings, background noise was not an issue, because the recordings were made in quiet places. Acoustic analysis of fundamental frequency was used to classify cries according to a set of canonical melody shape primitives. The frequency of occurrence of each melody shape was then plotted

across 12 different age groups ranging from 0 to 17 months of age. In earlier work, Varallyay Jr et al (2004) compared several acoustic features of cry vocalizations between normal and hard-of-hearing infants and found differences in fundamental frequency and dominant frequency.

On a related note, Garcia and Garcia (2003) developed an automatic recognition system for distinguishing between normal infant cries and the pathological cries of deaf infants using Mel-Frequency Cepstral Coefficients. The crying of infants ranging from 0 to 6 months of age was recorded in clinics by pediatricians, who also annotated the recordings at the end of each session. A total of 304 MFCC-related features were computed, submitted to PCA for dimensionality reduction, and finally applied to a feed-forward neural network. Up to 97.43% classification accuracy was achieved.

Like Garcia and Garcia (2003), Petroni et al (1995) used feed-forward neural networks for modeling acoustic attributes of infant cries. Here, the goal was to automate the classification of pain vs. non-pain infant cries. Inspired by the work of Waz-Hockert (1985) and Fuller (1991), their underlying motivation behind was developing an automatic recognition system for clinical settings to aid in "diagnosis of pathology or for identifying the potential of an infant at risk (Petroni, Malowany, Johnston, & Stevents, 1994)". Sixteen healthy 2-6 month old infants were recorded in a total of 230 cry episodes within a hospital setting. The cries were elicited by one of three stimulus situations: pain/distress, fear/startle, and anger/frustration. The latter two categories were grouped together as a single class, and the goal of the classification was to distinguish between pain/distress cries and everything else. Accuracies of up to 86.2% were achieved.

Papaeliou et al (2002) implemented acoustic modeling of emotion as a general mode of expression, in order to distinguish between emotive and communicative intent in infant vocalizations. Their motive in developing such a mechanism was to facilitate developmental study and diagnostic insight into communicative and emotional disorders, such as William and Down syndromes, as well as autism. Using this classifier to discriminate acoustically between the two modes of expression, their goal was to examine the hypothesis that infants' ability to vocally distinguish between emotions and communicative functions may serve as an index of their communicative competence later on (Bloom, 1998). According to Trevarthen (1990) and Bloom (1998), the extent to which

infants differentiate between emotional and communicative expression reflects how effectively they can regulate interpersonal communication. As a first step in examining this hypothesis, Papaeliou et al investigated whether infants' vocalizations can be differentiated acoustically between emotive and communicative functions in the second half of the first year of life. The vocal repertoire of infants between the ages of 7 to 11 months was recorded in the infants' homes, every 2 weeks in four 7-minute sessions of spontaneous play with their mothers. Like Speechome and several of the works listed above, they placed great value on ecologically valid observational conditions:

> "The home environment is considered more appropriate for obtaining representative samples of the infants' vocal repertoire than the unfamiliar laboratory environment."
> (Papaeliou et al., 2002)

Annotation to attribute emotional or communicative intent to each vocalization was done in separate interview sessions with the mothers, by having the mothers review the videotape recordings and identify what they felt their baby was expressing in each instance of videotaped interaction. In acoustic analysis of vocalizations, crying, laughing, vegetative (i.e., bodily) sounds, and child speech were excluded, as well as vocalizations that overlap with a caretaker's voice or external noise. A set of 21 acoustic features were computed for each vocalization and submitted to a Least Squares Minimum Distance classifier, using leave-one-out cross validation to evaluate recognition accuracy. An overall classification accuracy of 87.34% relative to mothers' interpretations was achieved.

1.1.2.3 Distinguishing Multiple Emotional States

As part of a study of infant vocal development, Scheiner et al (2002) built a classifier to distinguish acoustically between multiple emotional states. A comprehensive vocal repertoire was recorded of 7 healthy infants during the course of their first year of life, starting at the age of 7 to 10 weeks and ending at 53 to 58 weeks. Recordings were made by the parents themselves, in familiar surroundings, at intervals of 4 to 6 weeks. Each "session" took the form of a week-long assignment: parents were given a prescribed set of situations to record during the course of one week. The parents also served as the

annotators, indicating for each recorded situation the one of seven specific emotions they assumed their infant was expressing, such as joy, contentment, unease, anger, etc. Various acoustic features were computed for each vocalization, and a stepwise discriminant function was used to implement the classification. As part of the analysis, 11 call types were identified and used as additional discriminatory features to help with classification, including cry, coo/wail, babble, squeal, moan, etc. Significant differences in the acoustic structure of call types, and some emotional states within these call types, were found. In particular, discrimination was far better when grouping the more specific annotated emotional states into broad categories such as positive and negative. A longitudinal analysis also showed significant acoustic changes in some of the emotion-qualified (positive/negative) call types.

Another classifier similarly using vocal acoustics to distinguish emotional states of children was built for a very different purpose by Batliner et al (2006): to create social robots that respond appropriate to emotions in children's speech. In addition to the motivations they list – edutainment and child-robot communication – we note that there is an emerging research movement for developing social robots to assist in therapies for developmentally impaired children, particularly in autism (Scassellati, 2005). Autistic children have been observed to respond remarkably well to social robots, demonstrating "positive proto-social behaviors" that they rarely show otherwise (Scassellati, 2005). Social robots have been found to generate a high degree of motivation and engagement in autistic children, including those who are unlikely or unwilling to interact socially with human therapists (Scassellati, 2005). Thus, social robots can potentially be useful intermediaries between the autistic child and the caretaker or therapist, teaching autistic children about social interactions, and providing insights to better understand autistic behavior.

In Batliner's study, school-children aged 10-13 were individually given an AIBO robot to play with in a classroom within their school and instructed to talk to it "like they would talk to a friend" (Batliner et al., 2006). The play session was recorded and then annotated to capture the following emotional states in the child's speech: joyful, surprised, emphatic, helpless, touchy/irritated, angry, motherese, bored, reprimanding, and other non-neutral. Acoustic features of the child's utterances were used to train a classifier that combines several machine learning algorithms, such as neural networks, and support vector

machines. Various combinations were tested, and the best performing achieved a recognition rate of 63.5%.

1.2 Thesis Overview

In this section, we summarize the contributions of this thesis and provide a preview of the results.

1.2.1 Contributions

The contributions of this thesis are threefold:

First, we develop a methodology that is designed to introduce the child's perceived emotional state to the set of variables available for study in current and future Speechome datasets. With over 230,000 hours of recording, the scale of the Speechome corpus requires an automated approach. Our goal in this methodology is therefore to answer the question: ***How can we go about building a mechanism for automated recognition of a child's perceived emotional state in the Speechome corpus?*** We combine machine learning and data mining strategies to define a recipe for creating such a mechanism.

Second, using this methodology, we seek to derive a set of vocal acoustic correlates of the child's perceived emotional state, with which we can effectively model human perceptual heuristics, asking the question: ***How well can such a model simulate human perception of a child's emotional state?*** We model the child's perceived emotional state using acoustic features of the child's vocalizations and surrounding adult speech sampled from a child's 9-24 month developmental period. We evaluate the perceptual accuracy of models built using child-only, adult-only, and combined feature sets within the overall sampled dataset, as well as controlling for social situations, vocalization behaviors (such as crying, babble, speech, and laughing), individual caretakers, and developmental age. The perceptual accuracy of these models is given by the strength of the correlation between the acoustic vocal features chosen by the model and the child's perceived emotional state.

Third, as part of our comparative analysis we explore socio-behavioral, dyadic, and developmental patterns that emerge among these correlations. Increases in correlation

between a child's vocal expression and the child's perceived emotional state may indicate an increase in emotional expressiveness. Changes in correlation between surrounding adult speech and a child's perceived emotional state may suggest changes in child-caretaker intersubjectivity. In our exploratory analysis, we ask the following questions:

- ***To what degree does surrounding adult speech reflect the child's perceived emotional state?***

- ***Does the speech of certain caretakers correlate with the child's perceived emotional state more clearly than others?***

- ***Do any stronger correlations between perceived child emotions and adult speech emerge in social situations or when the child is crying, babbling, or laughing?***

- ***Are there any longitudinal trends in correlation that might indicate developmental changes in emotional expressiveness or intersubjectivity during the child's growth from 9 to 24 months of age?***

1.2.1.1 Developing a Methodology

Among the options available for modeling perceived emotional state and automating emotion recognition in a raw audio/video recording medium such as Speechome are vocal acoustics, facial expressions, and postural analysis. In this work, we focus on vocal expression as the first step towards a multimodal emotion recognition system, including both the child's vocalizations and surrounding adult speech in the modeling process.

The first step in modeling human perception of a child's emotional state is to gather examples of actual human perceptions with which to train the model. In machine learning terminology, these examples are called the *ground truth* of human perception by which the model is trained. As part of our methodology, we design, implement, and deploy an annotation interface, questionnaire, and sampling strategy for collecting ground truth from the Speechome recordings about the child's vocalizations and the perceived emotional

state that corresponds to each. We represent emotional state using two parameters – *valence* (i.e. Mood) and *arousal* (i.e. Energy) – that together form Schlosberg's two-dimensional space for characterizing affect (Schlosberg, 1952). Using the manually annotated child vocalization intervals, we implement a data mining solution to query the periods of adult speech from the Speechome corpus that occur in proximity to each vocalization. Computing a set of 82 acoustic features for each child vocalization and window of surrounding adult speech, we then apply Partial Least Squares regression to model the child's perceived emotional state given these features.

1.2.1.2 Building a Model for Perceiving Child Emotion

With Partial Least Squares regression as our modeling algorithm of choice, we apply 10-fold cross validation to evaluate the perceptual accuracy of these models. Our metric for these evaluations is adjusted R-squared, which is derived from the Mean Squared Error given by the cross-validation procedure and normalized based on sample size and number of modeled components. In addition to recognition success, adjusted R-squared is synonymously represented in analytical literature as a measure of goodness of fit, and therefore correlation, between predictor and response variables. In the context of our model, the acoustic features are the predictors, and the <Mood, Energy> pair representing the child's emotional state is the response.

We present our evaluation of perceptual accuracy as a comparative analysis between child-only, adult-only, and combined models across socio-behavioral, dyadic, and longitudinal subsets of the data. The purpose of this multidimensional analysis is to investigate methodological parameters for optimizing the design of an automated emotion recognition mechanism. For example, if perceptual accuracy is significantly higher in social situations, an emotion recognition mechanism would benefit from building a separate model for social situations that is applied conditionally only during social situations. Similarly, a particular caretaker's speech may happen to correlate exceptionally well with the child's perceived emotional state, improving the performance of an emotion recognition mechanism if applied conditionally in contexts involving that caretaker.

In addition to social situations, the socio-behavioral contexts that we use in this work vary according to different vocalization behaviors of the child: crying, laughing, babble, speech, bodily (e.g. sneezing, coughing) and other emoting. The Other Emoting category includes all other vocalizations, such as cooing, whining, fussiness, and squealing, which are for the most part amorphous and difficult to define and distinguish. We include these distinctions of vocal behavior in our comparisons in order to investigate whether vocalizations tied to specific emotions, such as crying and laughing, bear higher correlations with perceived emotional state. Also, we wish to control for possible differences in acoustic properties that may be inherent in certain vocalization types. In particular, vocal emotional expression may take on a different form in a general-purpose communication medium such as babble or speech than in a purely emotive vocalization such as crying, whining, squealing, laughing, or cooing. Also, we filter out bodily vocalizations such as sneezing and coughing, which carry no emotional significance, in order to evaluate the impact of the noise they might introduce in a model that includes all vocalizations regardless of contextual distinction.

Longitudinally, it is natural to expect that vocal emotional expression changes over time in the course of development. Research has found that the vocal tract anatomy experiences significant changes during early childhood, such as the condescensus of the larynx (Warlaumont et al., 2010). These physiological changes, together with cognitive growth, greater kinesthetic control of the vocal tract, and improvements in emotional self-regulation, all strongly suggest that vocal emotional expression changes developmentally as well. Building an emotion recognition mechanism to cover the entire 9 to 24 month period could abstract from the strengths of any correlations that may be specific to a particular stage of development. To investigate this hypothesis, we also build and evaluate models for each month's worth of data separately.

1.2.1.3 Exploratory Analysis

We plot adjusted R-squared across socio-behavioral, dyadic, and longitudinal subsets of the data and investigate whether there are any patterns that appear to be developmental, or conditional to a social situation, vocal behavior, or caretaker. In addition

to evaluating perceptual accuracy of each model, we interpret adjusted R-squared as an indicator of correlation between the child's perceived emotional state and the vocal expression of the child and caretakers. In the case of models that were built using adult-only feature sets, this correlation can be viewed as a measure of intersubjectivity, or empathy, reflecting caretakers' vocally expressed sensitivity to the child's emotional state.

1.2.2 Results preview

Our results show great potential for achieving robust vocal recognition of a child's perceived emotional state in the Speechome corpus, with child and combined models achieving overall adjusted R-squared values of 0.54 and 0.59, respectively. Interpretation of adjusted R-squared is given by a scale in which 0.2-0.12 is a small effect size, 0.13-0.25 is a medium effect size, and 0.26 or higher is a large effect size. Thus, we interpret the recognition performance of both child and combined models trained on the overall dataset to be consistent with an appreciably large effect size. As demonstrated by Figure 1-1(a), recognition performance is markedly consistent across socio-behavioral contexts.

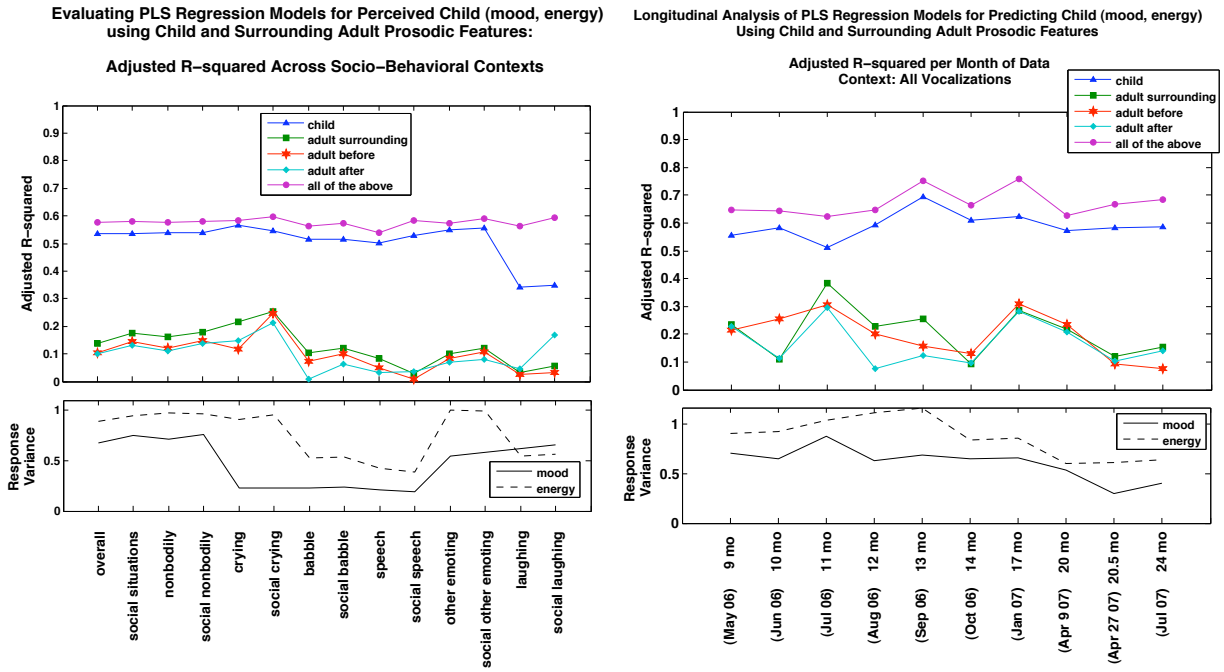


Figure 1-1. Previews of (a) Adjusted R-squared and Response Variance across Socio-Behavioral Contexts, for all time and all caretakers in aggregate. **(b)** Longitudinal Trends of Adjusted R-squared for All Vocalizations, regardless of Socio-Behavioral context.

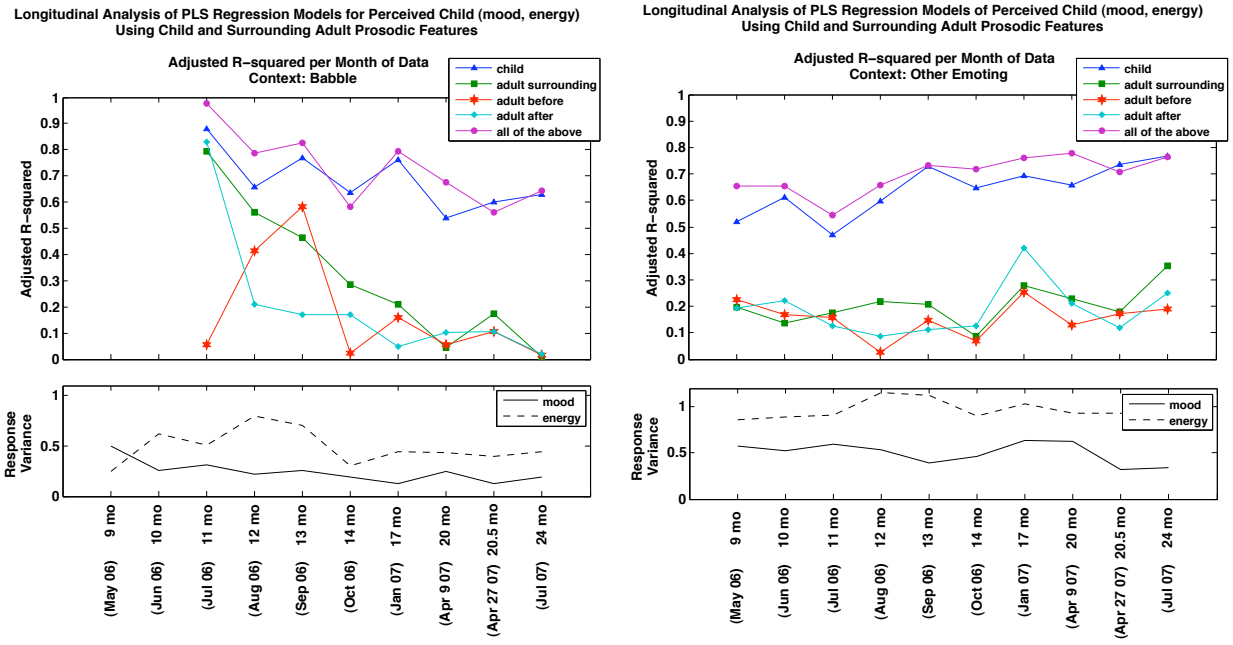


Figure 1-2. Previews of Longitudinal Trends of Adjusted R-squared for (a) Babble and (b) Other Emoting contexts. In Babble, we see a marked downward progression in correlation with the child's emotional state for adult speech surrounding the child's vocalizations, as well as a mild downward pattern in correlation of the child's vocal expression with the child's own emotional state. In Other Emoting, we see a mild upward progression in the child's emotional expressiveness over time that seems to counter the downward trend in Babble.

The high recognition performance described above corresponds to models built using data from the entire developmental time frame of 9-24 months in aggregate. In our longitudinal analysis, we observe the perceptual accuracy of month-specific models to be even higher: for the data covering all socio-behavioral contexts in aggregate, average adjusted R-squared was 0.67, ranging from 0.62 to 0.76. The longitudinal trend for this overall dataset, shown in Figure 1-1(b), does not show any progressive rise or fall in recognition accuracy over time. This suggests that in the general case, the correlation between perceived child emotion and both child and adult vocal expression is not subject to developmental change. However, we do observe striking longitudinal progressions when looking at trends for specific socio-behavioral contexts, notably Babble, and Other Emoting/Social Other Emoting, as shown in Figures 1-2(a, b). We discuss our developmental hypotheses regarding these progressions in Chapter 6.

1.3 Roadmap

The rest of this thesis is organized as follows. Chapter 2 provides some background about the Human Speechome Project – its origins, recording infrastructure, present data corpus, and future directions. In Chapter 3, we describe the data collection methodology – the implementation of the annotation interface, design and deployment of our sampling strategy and annotator questionnaire, our data mining solutions for minimizing annotation volume, and an evaluation of inter-annotator agreement. Our analysis methodology is detailed in Chapter 4, from post-processing of annotated data, to acoustic feature extraction, and ending with our use of Partial Least Squares regression: our rationale for choosing this method, the experimental design for our exploratory analysis, and an explanation of adjusted R-squared, our metric for evaluating recognition accuracy. Chapter 5 presents the results of the analysis, and we discuss their implications in Chapter 6.

Chapter 2

The Human Speechome Project

Because the goal of this thesis is to prepare the Speechome corpus for the study of emotional factors in child development, the Human Speechome Project is at the heart of our methodology. To put our work into context, this chapter describes the Human Speechome Project in greater detail – its origins, purpose, infrastructure, and speech database, focusing on aspects of each that are relevant to our work.

The Human Speechome Project (D. Roy et al., 2006), launched in August, 2005, was conceived with the complementary dual purpose of embedding the study of child language development in the context of the home, while capturing a record of the developmental process at an unprecedented longitudinal scale and density. The home of a family with a newborn child, who is now known to be typically developing, was equipped with cameras and microphones in every room, as well as a set of servers for data capture, all networked together to record and store the data. Over the course of three years, more than 230,000 hours of audio/video recordings were made of the child's daily waking life using this system.

This dense, longitudinal, naturalistic dataset is an innovation that promises fundamental advancement in understanding infant cognitive and linguistic development, and for understanding and treating developmental disorders that affect language and social interaction, such as autism. Its longitudinal density, characterized by an average of 9.5 hours of daily recording over the course of three years, enables the study of developmental research questions at variety of hierarchical time scales. The shape of developmental change can be very different at more fine-grained hourly or daily time scales than in a weekly or monthly big-picture perspective (Adolph et al., 2008), and depending on the research question, both can offer valuable insights about the same developmental process (Adolph et al., 2008; Noonan et al., 2004). For many developmental questions, the right sampling frequency is unknown a priori and must be determined empirically through trial

and error (Adolph et al., 2008). The ultra-dense nature of the Speechome corpus facilitates this kind of exploration and makes possible the discovery of developmental phenomena that occur at very fine-grained time scales.

The ecological validity of the Speechome corpus, characterized by recording of the child growing up within the natural context of the home setting and interacting with his caretakers in daily life situations, addresses a commonly expressed need among developmental researchers (Baranek, 1999; Bloom, 1973; Bruner, 1983; Nelson, 1974; D. Roy et al., 2006; Trevarthen, 1993). Traditional research studying child language development has typically asked parents to bring their children into a lab, or has sent researchers to videotape the child at home for just a couple hours per week. Bruner recognized the importance of context sensitivity in language acquisition processes and thus favored observation within the home setting (Bruner, 1983; D. Roy et al., 2006). Trevarthen echoes the same concerns regarding the study of socio-emotional development:

> We need to put observations of the communicative and emotional characteristics of infants, *when they are actively interacting with people and natural events and adaptively regulating their experiences*, in the context of experimental studies that have sought to measure different parts of the system in isolation.
>
> (Trevarthen, 1993)

A similar need for ecologically valid observation has been expressed regarding the psychotherapy of emotion dysregulation, as well as the study and treatment of developmental disorders such as autism (Baranek, 1999; Cole et al., 1994; Hayes et al., 2008; Hayes et al., 2004; Kientz et al., 2005). With ongoing initiatives to collect additional Speechome datasets across multiple neurotypical and autistic children, the Human Speechome Project has introduced new possibilities in meeting this need.

2.1 Data Capture

Each room and hallway of the participating family's home was equipped with omni-directional, mega-pixel resolution color digital cameras and highly sensitive boundary layer microphones, embedded in the ceiling as shown in Figure 2-1, totaling 11 cameras and 14 microphones. The cameras use a fisheye lens, which allows a complete panoramic view of a

room with little or no loss of data. Fisheye lenses are common in surveillance for this same reason. Motorized shutters conceal the cameras when not recording, swinging open when recording is turned on. The resulting view (Figure 2-2) provides wide coverage, at the expense of details such as facial expressions, which are lost due to the overhead perspective (D. Roy et al., 2006). The boundary layer microphones use the ceiling as a sound pickup surface, producing high quality speech recordings that reduce background noise, enabling reliable speech transcription. Their sensitivity can capture with clarity not only soft vocalizations even at a whisper level, but also subtle acoustic nuances of vocal expression more generally.

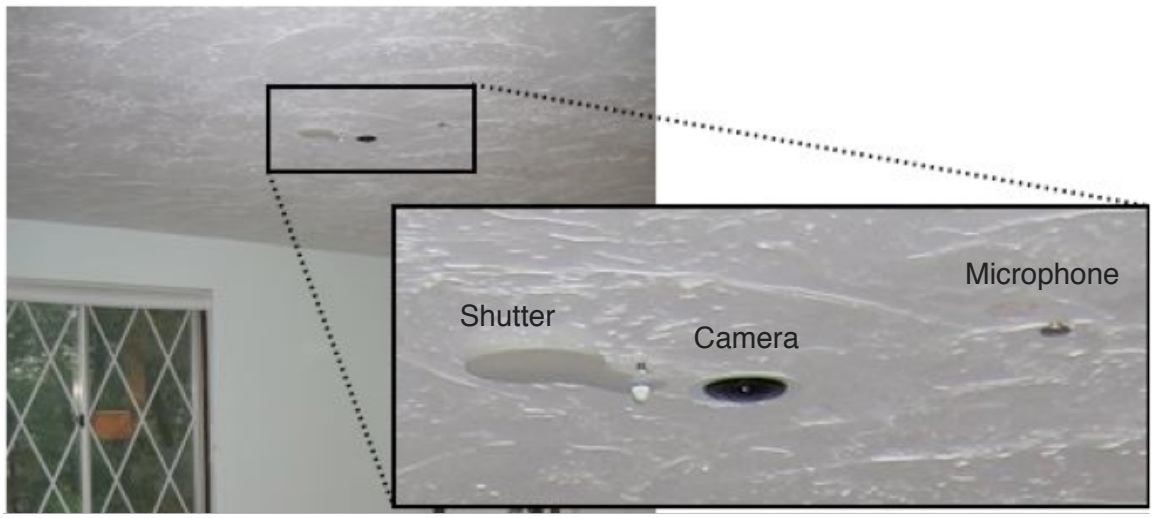


Figure 2-1. Speechome Camera and Microphone Embedded in the Ceiling of a Room.

The family was given full control of recording and privacy management; recording could be turned on and off at will using miniature wall-mounted touch displays, shown in Figure 2-3. The interface of these displays consisted of four virtual buttons: a video recording switch (camera icon), an audio recording switch (microphone icon), an "oops" button (exclamation point icon) for retroactively erasing recordings that the family may want off the record, and an "ooh" button (asterisk icon) for flagging notable events, such as the child's first steps. Eight such control devices were mounted throughout the house, next to the light switches, with each control mapping to cameras and microphones within a specific zone of the house.

Figure 2-2. Speechome Overhead Camera View

Figure 2-3. Speechome Recording Control Interface

The cameras and microphones are all network-enabled and were wired to a server infrastructure set up in the basement of the home. A specialized data capture application[1] was developed specifically for the Human Speechome Project, and was running continuously on these servers, sampling audio at greater than 48 KHz, capturing video at 14 frames per second whenever motion was detected (only one frame per second whenever motion was not detected), and performing real-time video compression. The server infrastructure and data capture application were designed to handle the high performance storage and computational processing involved in assembling an average of 9.6 hours of high quality audio/video recordings per day. With video compression, approximately 300 gigabytes of raw data accumulated each day (D. Roy et al., 2006).

Daily recording continued from birth through the age of three, adding up to over 230,000 hours of raw audio/video, and taking up at least one petabyte (1 million gigabytes) of storage (D. Roy, 2009; D. Roy et al., 2006). For the purpose of studying language development, post-processing, speech transcription, and generation of other new metadata has focused on the period during which the child ranges from 9 to 24 months of

[1] BlackEye, by Philip DeCamp.

age. This 488-day period consists of 4,260 hours of recording time across 444 days, capturing approximately 70-80% of the child's waking hours (D. Roy, 2009).

2.2 Speechome Database

To gather meaningful observations out of this vast sea of audio-video recording, it is necessary to first develop specialized data mining tools. A speech processing pipeline (D. Roy, 2009) consisting of an array of fully- and semi-automated technologies has been implemented to collect speech-related metadata from the raw Speechome audio/video recordings, for the purpose of creating a dataset for studying child language development. This dataset is commonly denoted as the *Speechome database*. Together, the raw media recordings and the Speechome database form the *Speechome corpus*. We describe in this section the portion of the speech processing pipeline and Speechome database that is relevant to our work in this thesis.

The main focus of this pipeline has been to obtain a comprehensive set of speech transcripts that capture all words heard and produced by the child from 9 to 24 months of age. Because of the naturalistic environment in which the Speechome recordings were made, the audio contains frequent instances of background noise, overlapping speakers, and spontaneous speaking styles with varying degrees of articulation, all of which present challenges to automatic speech recognition (D. Roy, 2009). With automatic speech recognition tests producing an error rate of well over 75% (D. Roy, 2009), a process of manual speech transcription became a necessary element in the speech processing pipeline.

The pipeline begins with fully automated mechanisms that extract the portions of the 14-channel audio stream that are most likely to contain speech and prepare them for streamlined manual speech transcription: these mechanisms are channel selection, speech detection, and speech segmentation (B. C. Roy & Roy, 2009; D. Roy et al., 2006). As described in (D. Roy, 2009), a Channel Selection module chooses the audio channel with the highest persistent energy as the most basic heuristic for narrowing down the audio field. Next, a Speech Detection module applies a boosted decision tree to classify 30ms frames of

audio as either speech or not speech. Of the frames identified as speech, a Speech Segmentation module stitches adjacent frames together and separates them at pauses to form speech segments. Speech segments are the basic audio units that are ultimately passed to manual speech transcription.

Prior to speech transcription, however, these speech segments are further filtered such that only child-available and child-produced speech is transcribed. This is done using metadata about child presence that has been previously collected using a combination of heuristic algorithms and manual video annotation. Using this information, the Child-Availability Filter retains for speech transcription only those speech segments that occur when the child is present in a given room.

Speech transcription itself is streamlined through the use of BlitzScribe, an interface that accelerates transcription time by a factor of 4 to 6 over other available transcription tools (B. C. Roy & Roy, 2009; D. Roy, 2009). BlitzScribe synchronizes playback to the transcriber's speech of transcription by internally applying the fully automated modules described above and passing the resulting speech segments to the transcriber, who then simply listens and types.

Finally, a speaker identification mechanism implemented by Miller (2009) for the Human Speechome Project uses a boosted decision-tree classifier to automatically identify speech segments as having been uttered by either the child, mother, father, nanny, or other. The classifier was trained using acoustic features[2] computed from a set of hand-labeled samples of each speaker's utterances. The *other* category includes houseguests, occasional nanny substitutes, sounds made by the child's toys, and voices heard over speakerphone, among other things.

The metadata and transcriptions generated via this semi-automated process of speech transcription collectively form the Speechome database. Speech transcription is ongoing at the time of this writing; at least 28% of the audio corpus covering the 9-24 month period has been transcribed, involving over 1,200 hours of transcribed speech segments and yielding over 3 million transcribed words (D. Roy, 2009).

---

[2] Mel-Frequency Cepstral Coefficients (MFCCs)

Chapter 3

Data Collection Methodology

The Speechome corpus, described in Chapter 2, promises an unprecedented perspective into child behavior and development within the context of the home and interactions with caretakers – a perspective that could yield insight and answer behavioral research questions on a rich hierarchy of time scales, from a second-by-second account of environmental factors that might suddenly cause a child to cry, to a developmental trajectory of emotional variability as it changes over the course of 24 months. Towards harnessing the benefits of such dense, longitudinal, naturalistic data for the study of child emotion, our goal in this thesis is to investigate acoustic correlates that can effectively model the child's perceived emotional state within the medium of the Speechome corpus.

The model that we propose to build determines the child's perceived emotional state using acoustic features of the child's vocalizations and surrounding adult speech within the Speechome audio. We represent emotional state according to Schlosberg's two-dimensional space for characterizing affect (Schlosberg, 1952). In our work specifically, it takes the form of a <Mood, Energy> pair, in which Mood and Energy are numerical variables that each range in value from 1 to 5. Mood represents the valence of the emotional state, which tells us whether, and to what extent, the child is sad (negative valence) or happy (positive valence). Energy represents the arousal in the emotional state – the degree to which the child is relaxed (low arousal) or excited (high arousal). We describe our methods of building such a model in Chapter 4 and evaluate its perceptual accuracy across longitudinal, dyadic, and socio-behavioral contexts in Chapter 5.

Before we can build such a model, however, our first challenge is to extract meaningful metadata from the raw Speechome recordings that would serve as the *ground truth* about the child's emotional state and vocalizations. Ground truth corresponds to human perception, the gold standard by which we wish to train our model to perceive the child's emotional state. Training a model involves giving it examples, or more precisely, sets of features that describe these examples, that are labeled with the human-perceived state.

Given enough examples, a well-trained model can accurately "perceive" the correct state, given a new set of input features. In order to collect these examples of human perception, manual human annotation is necessary. To set up large-scale manual annotation of the child's emotional state and vocalizations, we implemented a new annotation interface and hired a group of annotators to indicate their perceptions using this interface.

Annotators were provided with configured sets of Speechome audio/video clips that were sampled evenly throughout the child's 9-24 month developmental period. We describe our longitudinal sampling strategy in more detail in Section 3.2.3.3. Within each clip, annotators were asked to annotate time intervals for two types of events: child vocalizations, and adult speech or other noise that overlaps with the child vocalizations. For each vocalization, annotators rated the Mood and Energy of the child, each on a scale from 1 to 5, and answered questions pertaining to the behavioral nature of the vocalization (e.g. crying, laughing, babble), as well as whether it occurs in proximity to a social situation involving the child.

The rest of this chapter proceeds as follows. In section 3.1, we describe in detail our implementation of a highly configurable annotation interface and supporting infrastructure. Section 3.2 describes the particular configuration that we apply in deploying this interface for our specific research goals, as summarized above – including our questionnaire design, and our longitudinal sampling strategy in selecting clips for annotation. Section 3.3 describes the annotators and the annotator training process. Finally, in section 3.4 we present the results of our inter-annotator agreement analysis.

3.1 Infrastructure

Towards realizing a data collection process that is driven by manual human annotation, we[3] implemented a software infrastructure consisting of an annotation interface and a central back-end database, and set up a networked cluster of annotation machines to support multiple concurrent annotators. In our implementation, we take into

[3] In preference for the classic convention of academic writing, "we" and "our" always refers to work done by the author as part of this thesis.

account design considerations surrounding interface usability and portability, integration with persistent data storage that is transparent to the annotator, seamless administration of assignments, and ease of querying the data for analysis.

3.1.1 Interface Design and Implementation

The interface for browsing, playback, and annotation of the Speechome recordings was implemented using Java, with heavy use of its Swing/AWT libraries, and consists of three modular components: ***media browsing and playback***, ***time interval annotation***, and ***question & answer (Q&A) forms***. Although the annotation in this work primarily involves audio events, the interface was designed to support both audio and video annotation. In addition, the context afforded by multiple modalities makes video valuable for audio annotation, and vice versa. Thus, the ***media browsing and playback*** component implements a simultaneous, synchronized audio/video presentation of the Speechome recordings. The ***interval annotation component*** enables the user to specify the temporal range of specific event types observed during the course of media playback. ***Q&A forms*** present sets of questions for the annotator to answer about the recording and annotated intervals. The interval annotation component and Q&A forms are both easily configurable and can therefore be extended to arbitrary event types and question sets, as needed for other annotation goals. As will be described in Section 3.1.3, the database design supports this modularity.

3.1.1.1 Media Browsing and Playback

A pre-existing API[4] and utilities for playback of the raw, proprietary Speechome audio and video format files were available for programmatic use[5]. We developed our framework for browsing and retrieval of Speechome media clips using this Speechome playback API.

[4] An **Application Programming Interface** (API) is a set of software modules that have been developed to serve as a library for other programmers to use and apply for their own purposes, if the functionality of any of its modules fits their needs.

[5] Developed by Philip DeCamp.

The browsing and retrieval usage workflow first presents a list of clips, which we call a *playback browser list*, to the user, as shown in Appendix A, Figure 1. The user then browses the list by clicking on the desired clips, one at a time. The playback browser list takes as input a *clip configuration file* (CCF), which specifies a start date and time, end date and time, video channel number, and audio channel number for each clip, and builds the playback browser list at runtime based on the configuration specified by the input file. Times are specified to millisecond precision. Following is an arbitrary example of a specification of clip entries in a CCF:

2006-06-01 10:20:38.175 ~ 2006-06-01 10:20:39.345 ~ 23 ~ 3
2006-06-01 10:20:39.33 ~ 2006-06-01 10:20:42.54 ~ 23 ~ 3
2006-06-01 10:20:42.525 ~ 2006-06-01 10:20:44.58 ~ 23 ~ 3
2006-06-01 10:22:40.17 ~ 2006-06-01 10:22:42.6 ~ 23 ~ 3

When a user clicks on a particular clip in the playback browser list, the clip-specific *playback and annotation interface* (PAI) appears, which consists of a media playback module, the interval annotation interface, and Q&A forms, as indicated in Appendix A, Figure 2.

Features of the media playback module include a video playback viewer, a slider control that is synchronized with the timeline of the playback, a play/pause toggle button, and a fast-forward playback speed control.

3.1.1.2 Interval Annotation

Aligned below the media playback module, the interval annotation module consists of a set of *annotation tracks*, in which annotations made by the user appear, and an *audio waveform track*, as indicated in Appendix A, Figure 2. A vertical black line, or *tracer*, moves along the annotation and waveform tracks in sync with media playback, to support greater precision during annotation.

The interval annotation paradigm is determined by a mapping between event types, display colors, and keystrokes, which we call a *trackmap*. This mapping is completely configurable based on annotation needs (see Section 3.1.1.3), and is by no means limited to

the specific application presented in this thesis. The particular mapping configured for a runtime instance of the interface application is displayed for user reference at the upper right-hand area of the PAI, as shown in Appendix A, Figure 2.

Annotation Input Modalities

In the interval annotation module, one ***annotation track*** is created and displayed for each event type in the configured trackmap. When a user types one of the keys in the mapping while playback is on, an *interval annotation* begins to "evolve" in its corresponding annotation track, starting from the point in the playback timeline when the user pressed the key, in the color specified by the mapping. As playback continues, the interval annotation extends horizontally, in sync with the pace of playback, until the user presses that key again to toggle the annotation off.

This method of annotation, and several other features in the interval annotation module, were inspired by the elegant design of the VCode annotation interface (Hailpern & Hagedorn, 2008). VCode itself could not serve our purposes due to the nonstandard media formats in which the Speechome recordings are stored and the infeasibility of converting the multiple terabytes used in this work to standard formats. This incompatibility, initially considered a setback, necessitated the development of a new annotation interface – the interface being described in this chapter – which provided the opportunity to integrate the VCode-inspired annotation paradigm with a database infrastructure, and to include the Q&A forms (see Section 3.1.1.3) as part of the annotation process.

Developing a standalone annotation interface also enabled experimentation with other forms of user annotation input. The ***audio waveform track*** serves as a convenient, efficient alternative input modality, where the user can highlight the desired portion of the waveform and type the key that corresponds to the event being annotated. Highlighting the waveform is done by pressing the left mouse button at the desired start point, dragging to the desired end point, and releasing the left mouse button. The highlighted portion is displayed as a shaded region that extends as the mouse drags forward. When the mouse is released, playback is automatically initiated for just that highlighted portion.

If the user decides to annotate this highlighted time interval with one of the mapped event types, she types the corresponding key and the annotation interval appears in its correct track, for the exact interval denoted by the highlighted region. (The highlighting in the waveform track also automatically disappears at this time.) Otherwise, if the user decides not to make an annotation for the shaded region, the user can simply move on by highlighting another region, or by clicking the mouse anywhere in the interval annotation module. Benefits of this "highlight, listen, and type" annotation method include a more streamlined, pro-active annotation process and greater precision in annotating audio events due to the additional contextual guidance given by the waveform.

Object-oriented Hierarchy[6]

Each interval annotation is represented in the interface implementation as a distinct instance of a Java class, Annotation. Each track is implemented as an instance of the Track class, which acts as a container for Annotation objects of a particular color, corresponding to a particular keystroke. Annotation objects can be moved, deleted, or resized; we include these features to enable the annotator to adjust interval annotations for greater precision after they have been initiated, and to give the annotator an opportunity to re-evaluate the annotation after listening to the clip again.

An Annotation object is specified by a color, label, width, and horizontal start position within a Track. Vertical start position and height are constant for all Annotations. A unique *Annotation ID* is also generated for an Annotation object at creation time; this is the primary key that indexes the Annotation within the database (see Section 3.1.2). Upon creation, the Annotation object initiates an update to the database, which saves its start time, end time, and event label to the database schema, along with its unique Annotation ID

[6] For the non-technical reader, Object-oriented programming (OOP) is a software development methodology that organizes implementation into a set of "Objects," or "Classes." Each Object type is defined by the programmer to contain of a set of functions and state variables; these variables can be other Objects with different functionalities. Certain Objects might inherit functionality from other Objects. These containment and inheritance relationships among different Object types form a hierarchical, object-oriented implementation. Java, our language of choice, is an Object-oriented programming language, which is why our description includes references to "Objects" and "Classes."

and metadata to identify the user who made the annotation and the clip in which this annotation was made. Section 3.1.2 describes this database schema in greater detail. Transformation functions to map between horizontal track image space (i.e. pixel space) and playback timeline space are implemented as part of the **Track** object functionality. During the database update process, these transformations are applied to the horizontal start and end coordinates of an Annotation object to obtain the timestamps that define the annotation interval.

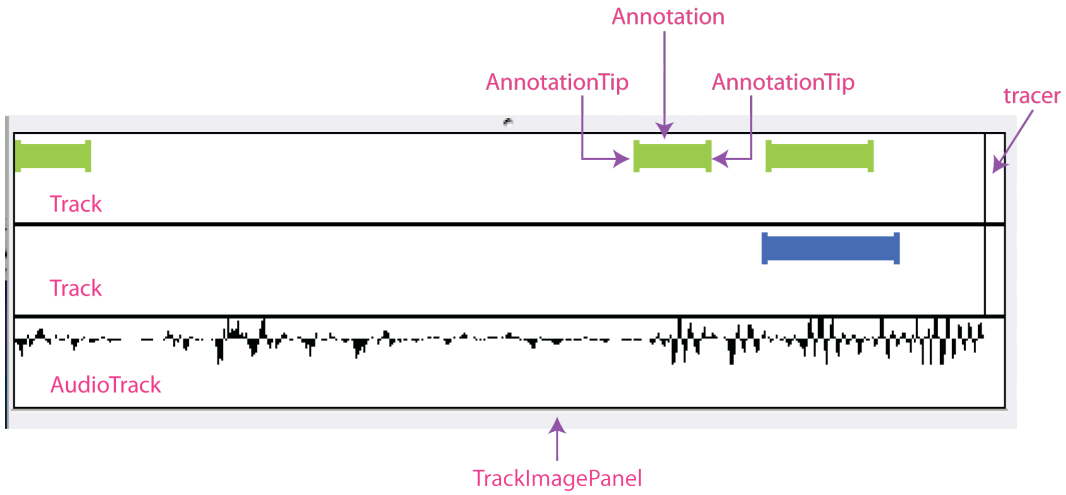


Figure 3-1. Annotation Interface Components. The annotation interface consists of a set of Track objects, one of which is a specialized AudioTrack. The regular Track objects are containers for Annotation objects – each Track holds a different event type. An Annotation object includes two AnnotationTips, one at each end, which enable the user to resize the Annotation object. The AudioTrack presents the waveform of the audio and implements the "highlight, listen, and type" annotation method.

The interval annotation module panel, **TrackImagePanel**, is a container for multiple **Track** objects, as well as a single **AudioTrack** object, which is a subclass of **Track**. The **AudioTrack** object constructs a smoothed waveform image from the audio clip and implements the "highlight, listen, and type" annotation input modality described above. The waveform image is computed by extracting amplitudes from the raw bepcm format of the audio for the presented clip (sampling rate 48000 Hz) and smoothing using a three-point averaging filter.

This hierarchical, object-oriented design (Figure 3-1) provides a modular platform for features such as resizing, moving and deleting **Annotation** objects within a **Track**. Each of these operations immediately causes the **Annotation** object to update its entry in the

database accordingly. For simplicity and usability, an additional element of design was needed to make the resizing and moving features as intuitive as possible for the user. Both moving and resizing an **Annotation** lend themselves most naturally to the ubiquitous left-click-and-drag mouse input paradigm. However, because moving and resizing are two separate functions, this presents the problem of how to enable the user to apply the same input modality for two different features.

Fortunately, there is a difference in the position where the user would tend to click an object in order to move it, as opposed to resizing it. For example, this difference is apparent in the design of windowed operating system interfaces, as shown in Figure 3-2, which harnesses the natural tendency to initiate a move of a window or any onscreen element by clicking somewhere in the middle of the operable region. Similarly, the intention to resize tends to focus the user's attention to the edges of a window.

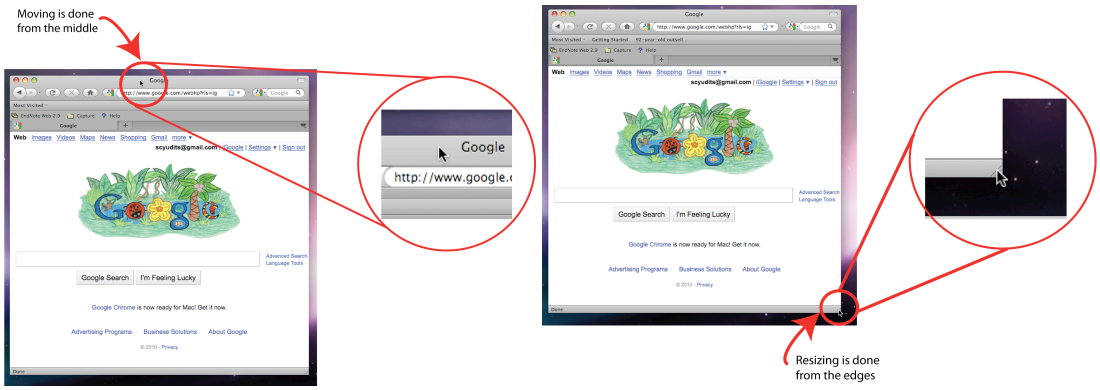


Figure 3-2. Moving vs. Resizing in a Windowed Operating System.

Whether the design of move and resize operations in windowed operating systems has been inspired by natural user tendencies, or users have been ingrained into these tendencies after decades of working with such operating systems, this usage model is intuitive for users. We therefore felt that it should be both preserved and harnessed in the usage model for moving and resizing **Annotation** objects.

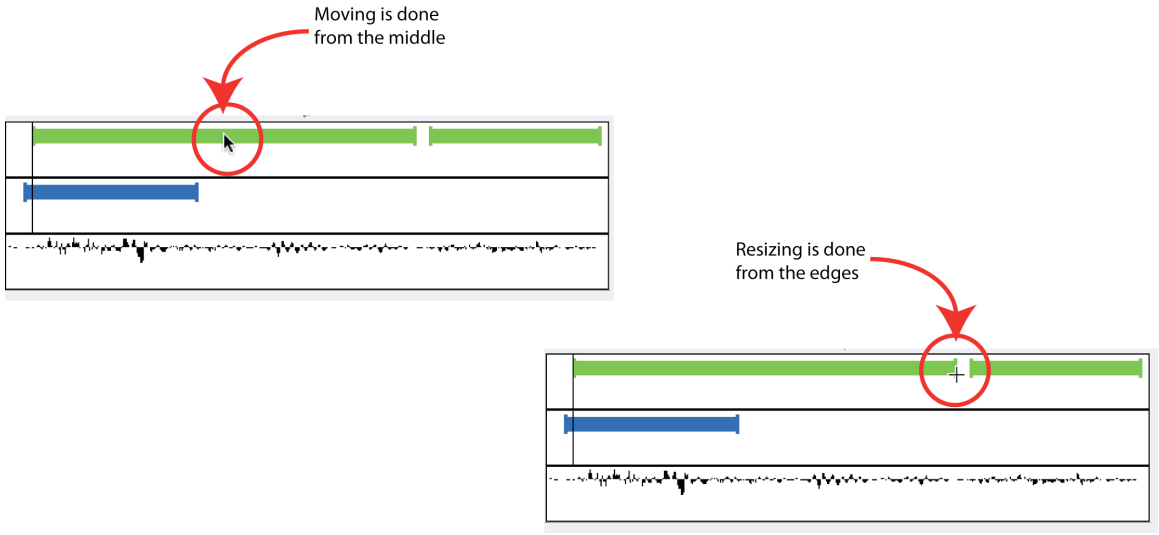


Figure 3-3. Moving vs. Resizing Annotations.

To implement this, we developed a new object type, called an **AnnotationTip**, represented graphically as a thin vertically oriented rectangle (Figure 3-1). Each **Annotation** object contains two **AnnotationTip**s of the same color – one located at the start position of the **Annotation** object, and the other at the end position. Resizing is done by left-clicking and dragging an **AnnotationTip**, while moving is done by left-clicking anywhere on the **Annotation** object in the region between the **AnnotationTip**s and dragging to the desired new location. When the mouse moves over an **AnnotationTip**, the shape of the cursor changes from an arrow into a pair of perpendicularly crossed lines, which helps guide the user to the appropriate usage and region distinctions.

3.1.1.3 Question & Answer Forms

In the course of usage workflow, the annotator also encounters the question & answer (Q&A) forms feature, which presents to the annotator a set of questions to be answered about a clip. The Q&A forms feature implements a mechanism for collecting metadata about the clip and each interval annotation as needed for the research question at hand. Its highly configurable design enables the researcher to develop a questionnaire that best serves the research question at hand, and to fluidly modify it throughout the research

process by simply specifying a different *question configuration file (QCF)*. Figure 3-4 shows the basic QCF format. For the actual QCF used in this work, please refer to Appendix B. The resulting interface display is shown in Appendix A, Figures 2 and 3.

```
Multichoice;Label1;Choice1^Choice2^Choice3
Scale;Label2;Choice1@1 - Choice2@2 - Choice3@3 – Choice4@4 - Choice5@5
Multichoice_Ann_v;Label3;Choice1^Choice2
Scale_Ann_v;Label4; Choice1@1 - Choice2@2 - Choice3@3
Scale_Ann_s;Label5; Low@1 - 2@2 - Medium@3 - 4@4 - High@5;3
Trackmap;Label6;v,Event1,Green,true ~ a,Event2,Blue,true ~ s,Event3,Orange,true
```

Figure 3-4. Question Configuration File (QCF) Format. Each line configures a new question to be presented to the annotator through the annotation interface's Q&A forms feature. Lines 1 and 3 define Multichoice questions, and lines 2, 4, and 5 define Scale questions. Lines 1 and 2 specify clip-level questions. Lines 3 and 4 specify annotation-level questions that appear only when annotations are created with the keystroke ***v***. Line 5 specifies an annotation level question of type Scale that will appear only for annotations created with keystroke ***s***, with default value preset to **3** in its scale slider. Line 6 specifies the trackmap for interval annotation, which defines three event types, corresponding to keystrokes ***v***, ***a***, and ***s***.

To summarize, each line of the QCF defines a separate question, using four fields – type, label, choices, and an optional default value. There are three types of questions that are supported by this interface: *Multichoice*, *Scale,* and *Trackmap*. A Multichoice question presents a list of choices, from which the user must select only one. A Scale question presents a range to the user in the form of a slider, with ticks at integer numerical values that correspond to question-specific nominal values, and asks the user to indicate a rating in the slider relative to the presented range.

The Trackmap entry is not a question, per se, but instead defines the mapping between event types, display colors, and keystrokes that is put into effect for interval annotation. Only one Trackmap is recognized per runtime instance of the interface. Each Trackmap "choice" maps a keystroke to an event type and a color. The interval annotation module (see Section 3.1.1.2) depends on this entry in the QCF to configure the number of Tracks and the keystrokes and colors that correspond to each Track. If the Trackmap specification is omitted from the QCF, no TrackImagePanel is displayed. In this case, no interval annotation is possible, and the Playback and Annotation Interface (PAI) can only be used as a clip-browsing interface.

The question **type** specifies not only whether it is a Multichoice, Scale, or Trackmap question, but also whether it is to be presented to the user at the clip level or at the annotation level. The latter is only applicable to Multichoice and Scale questions. When presented at the clip level, questions appear on the right hand panel of the PAI, below the Trackmap (see Appendix A, Figure 2). Annotation-level questions appear in a pop-up dialog window whenever the user creates a new annotation of that type, as shown in Appendix A, Figure 3.

The type field for annotation-level questions must additionally specify the annotation event type to which it corresponds. This is indicated using the key character that maps to that event in the Trackmap. For example, question type Multichoice_Ann_a defines a multichoice question that appears only when annotations are created with the keystroke ***a***, while a Multichoice_Ann_b question would only appear for annotations created with the keystroke ***b***. The particular keystroke characters used in this specification must correspond to one of the event type mappings in the Trackmap configuration described above.

The question **label** is the text of the question being asked, and the **choices** are specified according to the format that corresponds to the question type, as shown in Figure 3-4. Optionally, a **default value** can also be specified for a question, so that this choice appears preselected in the Q&A form before the user makes a selection. Configuring a default value is of particular use for scale questions, because they automatically come with a preset value at the beginning of the slider when a default value is not specified. In many cases, the beginning (i.e. extreme left) value of the slider represents an extreme condition or rating; if the slider comes preset to an extreme value, this could introduce bias into the mind of the user that subconsciously prefers values closer to that extreme. Thus, for certain kinds of questions, it can be more desirable to have the default value set at the middle of the slider, so that user bias is centered around a neutral or medium rating.

3.1.2 Administration Interface

We also implemented an administrative tool for use by the researcher, to facilitate distribution of assignments, creation of new user accounts, and checking the status of

assignments for each annotator. The interface, shown in Figure 3-5, presents a list of clip configuration files (CCFs) and question configuration files (QCFs) to the administrator. An assignment is created for a selected user by choosing one CCF and one QCF from their respective lists and pressing "Submit." These lists are constructed from the contents of two dedicated directories, <clipDir> and <qDir>, which separately hold CCFs and QCFs, respectively. The paths to <clipDir> and <qDir> are specified as input parameters. The administrative tool also includes functionality for creating new users and viewing the completion status of each user's assignments.

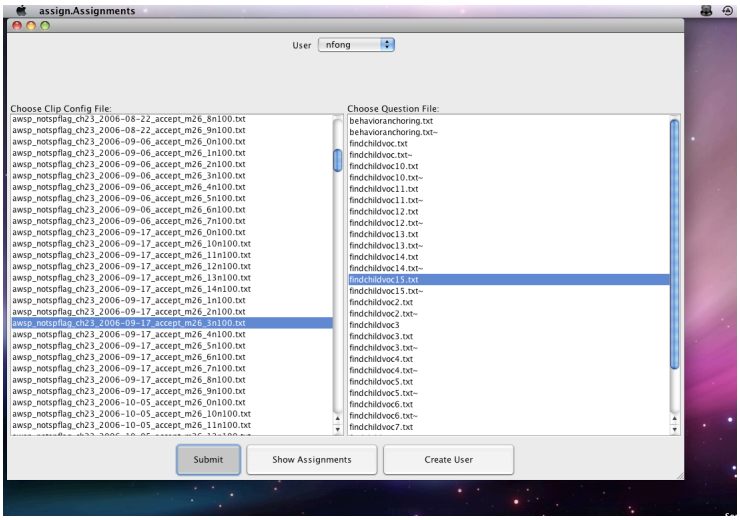


Figure 3-5. Administration Interface. Assignments are made by choosing an annotator from the User menu at the top, a Clip Configuration File (CCF) from the list on the left, and a Question Configuration File (QCF) from the list on the right, and clicking "Submit." New user accounts are created by clicking "Create User." Completion status of assignments can be viewed per annotator, under "Show Assignments."

3.1.3 Database Design & Usage Workflow

The annotation interface and the administration interface both communicate with a database that was set up to hold the annotation data and questionnaire answers. The database runs on a PostgreSQL v8.3 database management system (DBMS) platform, and the interfaces connect to it through SSL, using the Java JDBC libraries. Figure 3-6 summarizes the schema design. The rest of this section describes the details of this design

in context, by stepping through the usage workflow, from administration of assignments, to the creation of new annotations and submission of completed clips by the annotator.

3.1.3.1 Administration Workflow

When the administrator creates a new assignment using the administration tool, the selected CCF and QCF are imported into the database. CCFs, also called *clip lists*, are stored in the **clip_lists** table, indexed by a unique **clip_list_id**. A clip list is specified by its file name, which is stored in the **description** field, and the file contents, stored as the **clip_config_file** field. QCFs are stored in a similar way in the **anntypes** table, indexed by a unique **anntype_id** and specified by a **description** (the name of the file), and **qfile** (the contents of the file).

If **clip_lists** does not already have an entry that matches the contents and file name of the CCF being submitted, then this CCF is added as a new entry. Next, its contents are parsed, each line specifying an individual clip, as described in Section 3.1.1.1. For each of these clips, a new entry is created in the **clips** table. The **clips** table stores global clip specifications independently of any particular assignment, including the start and end times of the clip (**clip_start**, **clip_end**), the audio and video channel numbers (**audio_channel**, **video_channel**), the **clip_list_id** from which it came, and a unique **clip_id** index.

Similarly, the QCF submitted to the database through the administration tool as part of the creation of a new assignment is first checked against the existing entries of the **anntypes** table. If no matching file name and contents are found, then it is added to the table and parsed to extract and index the individual questions that are specified by the QCF. These individual questions are stored in the **questions** table, specified by the **question_label**, **qtype_name** (the question type, as described in Section 3.1.1.3), the **anntype_id** of the QCF from which it came, and a unique **question_id** index. The choices for each question are indexed in the **question_choices** table, which maps each **choice_label** and unique **choice_id** to its corresponding **question_id**.

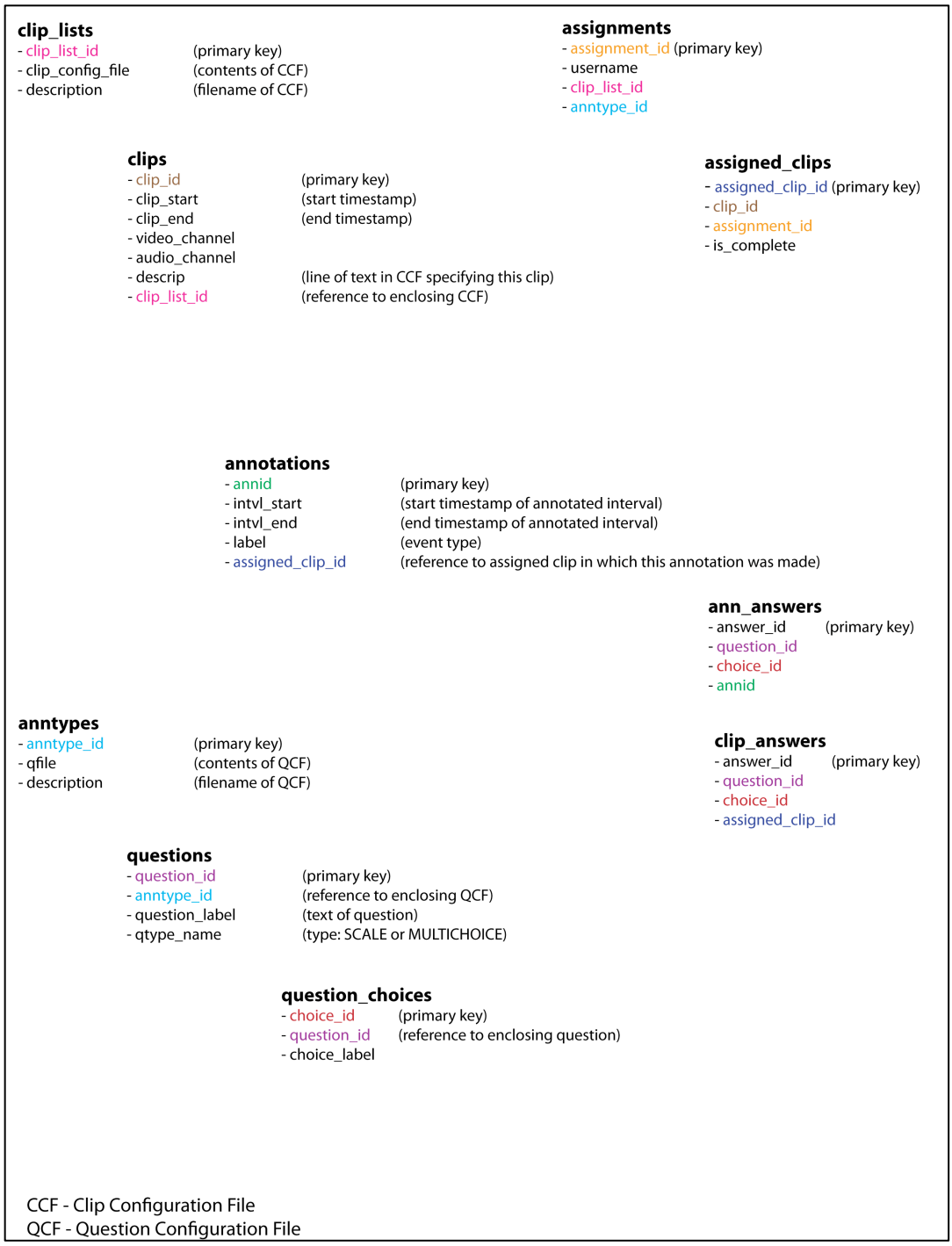


Figure 3-6. Schema Design. Tables are listed in bold. Fields are color coded to indicate cross-references that join two tables together via primary keys. For example, because assignments are defined by a CCF/QCF pair, each entry in the **assignments** table is specified in part by a reference to an entry in the clip_lists table, indexed by its primary key clip_list_id, and a reference to an entry in the anntypes table, indexed by its primary key, anntype_id.

An entry is then added to the **assignments** table to create the assignment; it is specified by the **clip_list_id** and **anntype_id** of its corresponding CCF and QCF, respectively, as well as the **username** to which it is assigned and a unique **assignment_id**. Next, each clip that belongs to the CCF indicated by this assignment's **clip_list_id** is imported into the **assigned_clips** table, with a unique **assigned_clip_id**, where it is mapped to its corresponding global **clip_id** indexed in the **clips** table and its **assignment_id**. A fourth field, **is_complete**, is initially set to *false*; this value becomes *true* when the annotator denoted by **username** submits her annotations and Q&A for this clip.

3.1.3.2 Annotation Workflow

When the annotator logs into the annotation interface, the application retrieves her active assignments from the **assignments** table and computes the percentage of assigned clips that have been completed within each assignment. These assignments are displayed as a list, as shown in Figure 3-7, from which the annotator can select one assignment at a time. A status bar indicates the proportion complete for each assignment. In the administration tool, the administrator can mark assignments as deactivated so that they no longer appear in this list. Upon selecting an assignment from her list, the annotator receives the playback browser list described in Section 3.1.1.1 and shown in Appendix A, Figure 1. Selecting a clip from the list brings up the playback and annotation interface (PAI), as shown in Appendix A, Figure 2.

Whenever a new annotation is created by the annotator in the PAI, a new entry is created in the **annotations** table, specifying the start and end times of the annotated interval (**intvl_start**, **intvl_end**), the event label of the annotation (**label**), and the **assigned_clip_id** that defines the context of this annotation. The **assigned_clip_id** in turn maps to its **assignment_id** and the **username**, thus allowing a full accounting of which annotations were made by each user, in each assignment. Each annotation receives its own unique identifier, **annid**.

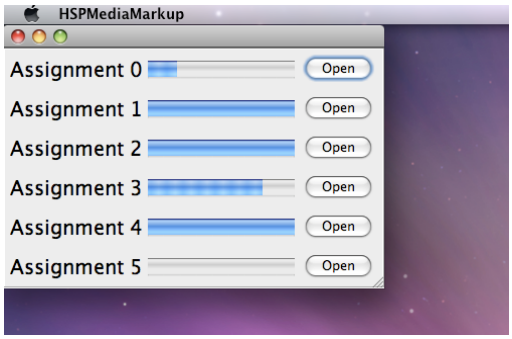


Figure 3-7. Annotator's List of Assignments. Status bars indicate the percentage of clips that have been completed in each assignment. To open an assignment, the annotator clicks its corresponding "Open" button.

Answers to Q&A forms are stored in the **clip_answers** and **ann_answers** tables for clip-level and annotation-level questions, respectively. Both tables map the **question_id** to the selected **choice_id**, and indexes each answer with a unique **answer_id**. The **clip_answers** table also maps each answer entry to its corresponding **assigned_clip_id**, while the **ann_answers** table maps each answer entry to its corresponding **annid**. When the annotator is finished making annotations and answering questions for a clip, she presses the "Finished, submit clip for review" button, which closes the PAI for the current clip and changes the value of **is_complete** to true in the **assigned_clips** table. Henceforth, this clip is listed in the playback browser list with a marker indicating its completed status, as indicated in Appendix A, Figure 1.

The annotator can also return to a previously annotated clip or specific annotation to review or change answers and interval placements. Whenever the application opens the PAI for a clip or the Q&A window for an annotation, it always checks the database first for any annotations and answers that this user may have already made for that assigned clip or specific annotation in the past, and imports them into the interface display.

3.2 Applied Input Configuration

With a modular, highly configurable annotation infrastructure in place, we designed a specific trackmap, questionnaire, and Speechome dataset sampling strategy to address the research questions in this work, and applied them as input configuration parameters to the

annotation interface. This section describes the design choices that we made in this configuration.

3.2.1 Trackmap

Annotators were asked to annotate two types of events: child vocalizations, and adult speech or other noise that overlapped with the child vocalizations. Appendix A, Figure 2 indicates this trackmap configuration, as seen by the annotator. Child vocalization events were mapped to the "v" key and displayed in green in the annotation track. Overlapping adult speech/noise events were mapped to the "a" key and displayed in blue.

3.2.2 Questionnaire

The questionnaire design includes both clip-level and annotation-level questions. At the clip level, annotators are asked the following questions:

1. **Does this clip include child vocalizations?**

Options are **yes** or **no**. If the answer is **no**, then the annotator is free to move on to the next clip.

2. **Are there adults talking or other noise during any part of the child's vocalizations?**

Options are **yes** or **no**. If the answer is **yes**, then the time intervals of the overlapping noise must be annotated.

3. **Is there any activity in this clip that might suggest a social situation involving the child and a caretaker? (See instruction sheet for examples.)**

Options are **yes** or **no**. This question was added to enable comparative study of the child's perceived emotional state between social and non-social situations.

Each clip could have multiple child vocalizations. Because an annotator may wish to answer vocalization-specific questions differently for each individual vocalization, we also present to the annotator a set of annotation-level questions, which appear in a new window whenever the annotator creates a new child vocalization annotation. For each child vocalization, annotators are asked the following questions:

1. **Which of the following best represents the nature of the vocalization?**

Options are **crying**, **laughing**, **other emoting**, **bodily (e.g. sneeze, cough)**, **babble**, **speech**. This question was included to serve several purposes. First, it enables us to separately study vocalizations that are clearly crying or laughing. Secondly, the **bodily (e.g. sneeze, cough)** option enables us to weed out vegetative sounds that by definition carry no emotional message. Finally, because **babble/speech** has clear syllabic structure that pure emoting (which includes **crying**, **laughing**, and **other emoting**) does not generally have, the acoustic vocal correlates for the same perceived emotional state may be different between these two classes of vocalizations, simply due to fundamental acoustic differences between them. Thus, it may be useful to build separate models depending on the nature of the vocalization.

2. **Please rate the energy of this vocalization. If it varies within the vocalization, rate its maximum energy: (1 = Lowest energy, 5 = Highest energy).**

Options are integers on a scale from 1 to 5, with 1 labeled **Low**, 3 labeled **Medium**, and 5 labeled **High**. This question is intended to capture *arousal*, one axis of Schlosberg's two-dimensional space for characterizing affect (Schlosberg, 1952). Ratings of 1 could include soft cooing or a whining sigh or a raspberry (Scheiner et al., 2002). These ratings often occur when the child is sleepy or very relaxed. Ratings of 2 might include absent-minded babble that is ongoing while the child is

occupied with something else. Regular babble epitomizes a 3 rating, although some instances of animated whining and cooing may also fall into this category. Many instances of crying, short of outright screaming, would fall into the category of 4, along with animated babble. Any kind of screaming or shouting, whether due to crying or excitement, would correspond to an energy label of 5.

3. **Please rate the child's mood on the following scale: (1 = Most negative, 5 = Most positive).**

Options are integers on a scale from 1 to 5, with 1 labeled **Negative**, 3 labeled **Neutral**, and 5 labeled **Positive**. This question captures *valence*, the second axis of Schlosberg's two-dimensional space for characterizing affect (Schlosberg, 1952). A value of 1, or Most Negative, represents a very sad mood, such as the child crying. A valence rating of 2 includes instances of slight whining or fussiness. Babble in which there is no obvious positive or negative mood is an example of a rating of 3. Cooing and pleasant "happy-baby" sounds fall under the category of 4. A 5 valence includes expressions of excitement, delight, and amusement that characterize an especially happy mood.

4. **How clearly do your answers above describe the vocalization?**

Options are integers on a scale from 1 to 5, with 1 labeled **No choices fit**, 3 labeled **Multiple choices fit equally**, and 5 labeled **Clear fit**. This question was intended primarily to help explain and mediate disagreements between annotators.

3.2.3 Input Dataset

To derive the set of clips applied as the input dataset for annotation, we put together a strategic combination of existing Speechome metadata: automatic speech detection (B. C. Roy & Roy, 2009), child presence video annotations (B. C. Roy, 2007), and speaker identification labels (Miller, 2009). In addition, we designed a longitudinal sampling

strategy to allow exploration of developmental trajectories in relation to the research questions in this work.

3.2.3.1 Existing Relevant Speechome Metadata

As described in Chapter 2, prior work by Brandon Roy (B. C. Roy, 2007; B. C. Roy & Roy, 2009) developed a mechanism for automatic detection of speech and applied it to the raw Speechome audio data. This effort produced a database of speech *segments* – a set of sound clips that includes all sound produced by human voice in the Speechome corpus, segmented at pauses and syllabic junctures occurring within the flow of human vocal expression. Each segment consists of a time interval specification and an association with a particular room of the house, based on the audio or video channel that recorded it. Each channel corresponds to an individual microphone (audio) or camera (video) device that was installed in a particular room of the house.

While comprehensive, the set of segments produced by the automatic speech detection mechanism contains many false positives – segments that only contain non-speech sounds that were not produced by the human voice, such as sirens in the distance, music, noise from laundry, and toy sounds. For this reason, the manual speech transcription process developed by Roy (2007) includes an option for the transcriber to mark a segment as "Not Speech". These flags are also stored in the Speechome database, and were applied in this work to filter out the non-speech segments.

Roy (2007) also previously implemented and deployed a video annotation process to track the presence of the child within the home. This annotation data also produced segments; in this context, each segment represents a time interval during which the child was present in a particular room.

An automated machine learning algorithm implemented by Miller (2009) labeled all speech segments in the Speechome database with speaker identification. Each label was assigned along with a confidence rating – a fractional value between 0 (lowest confidence) and 1 (highest confidence). This algorithm assigned five different labels, corresponding to the father, the mother, the primary nanny, the child, and other. The *other* category includes

houseguests, occasional nanny substitutes, sounds made by the child's toys, and voices heard over speakerphone, among other things.

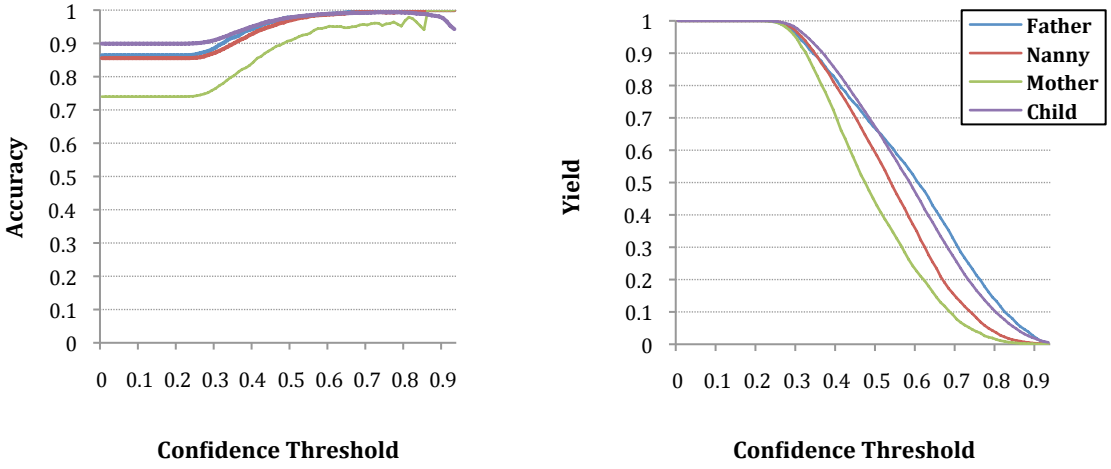


Figure 3-8. Accuracy and Yield of Miller's Speaker Identification Algorithm.
(data credit: (Miller, 2009))

Figure 3-8 shows the accuracy and yield of Miller's speaker identification algorithm for different confidence thresholds and speakers (Miller, 2009). The confidence threshold in these graphs specifies the minimum confidence value required to accept a speaker ID label assigned by the algorithm; any segments labeled with confidence value less than this threshold are deemed inconclusive with respect to specific speaker identification. Depending on the chosen confidence threshold and the particular speaker, the accuracy of these speaker classifications ranges from 73.9% to 100%, while carrying a tradeoff of decreasing yield with increasing accuracy. Accuracy for the mother (73.9%) at threshold = 0 is significantly lower than for the father (86.4%), nanny (85.5%), and child (89.9%), and this relative pattern continues even as all the accuracies increase at higher thresholds.

3.2.3.2 Constructing the Input Dataset

Because the annotation in this work focuses on isolating and characterizing child vocalizations, our goal in constructing the input dataset for annotation was to retrieve all audio that could possibly contain child vocalizations, while minimizing the number of

irrelevant clips that the annotators would need to sort through. Unlike video, which can be fast-forwarded to some degree without loss of meaning, audio must be heard at its naturally recorded rate, in order to maximize perceptual accuracy of the child's emotional state. Thus, it is critical to filter out as much irrelevant audio as possible from the annotation dataset.

To this end, we began constructing the annotation dataset by focusing on a single room in the house – the living room; only segments associated with this room are used. The rationale for limiting the current study to one room, despite the availability of multiple rooms in the Speechome corpus, is twofold. First, this focus enables us to explore the methodology in this work longitudinally through the complete 9 to 24 month period while keeping the amount of audio to be annotated within a manageable volume. Second, a single-room developmental observation model is consistent with future plans for the Speechome Recorder, described in Chapter 6. A compact, portable version of the original Speechome recording setup, the Speechome Recorder has been developed to enable large-scale deployment of the Speechome recording technology. A Speechome Recorder will be placed in a single room in each participating child's home. This room will be chosen by the family based on where both the child and caretakers spend a large portion of the day in play and other daily activity. Thus, the methodology developed and examined in our work, and the results of our analysis, based on data from a single room, would be directly applicable to the new corpora forthcoming from the Speechome Recorders.

We initially constructed the target dataset by taking the intersection between the speech segments and the segments indicating the presence of the child in the room. Later, the speaker identification labels and associated confidences were used to filter out segments likely to contain no child vocalizations. A parametric analysis, described in detail in Section 3.2.3.4, enabled us to choose an optimal configuration of confidence thresholds for each adult speaker that minimizes the number of child vocalizations that are filtered out, while also minimizing the number of irrelevant segments – purely adult speech or other noise – that are left in the dataset. Finally, the configuration of confidence thresholds determined to be optimal in this analysis was applied to the rest of the dataset to render the filtering.

3.2.3.3 Longitudinal Sampling Strategy

In the Speechome dataset, a day holds particular relevance as a fundamental unit of segmentation in the flow of observational data, because activities within a day are self-contained and do not carry over into the next day. In contrast, shorter time units such as hours and minutes are relatively arbitrary discrete boundaries when it comes to most human activities, even the fairly regular ones such as mealtimes, as they arise and dissipate in a continuous manner within the flow of daily life. This is even more the case when it comes to the spontaneous, volatile, dynamic realm of a child's emotions. We therefore used the "day" as our sampling unit.

At least two days per month were sampled on a monthly basis from 9-15 months, and one day per month every 3 months from 15 to 24 months. Each day's worth of segments, derived according to the method described in Section 3.2.3.2, was annotated in full. The chosen dates are listed below:

05/16/2006	07/02/2006	10/05/2006
05/21/2006	07/10/2006	10/19/2006
06/01/2006	08/07/2006	01/04/2007
06/08/2006	08/22/2006	04/09/2007
06/11/2006	09/06/2006	04/27/2007
06/24/2006	09/17/2006	07/06/2007

Where two or more days per month were chosen for annotation, at least one day was chosen near the beginning of the month, and another day near the end of the month, to spread out the sampling as evenly as possible. Due to the real-world nature of the Speechome recordings and the non-sequential sampling strategy of speech transcription, the spacing between days varies, being partly conditional on how much recording data is available for any given day, and whether the full set of metadata described in Section 3.2.3.1 is available yet for that day. (For example, there are no speech segments or other metadata available at all for the entire second half of July, 2006 at the time of this writing.) Some days had very little recording to begin with, and certain other days had corruptions, missing audio or video, or background noise (e.g. due to laundry) that could potentially be problematic in audio analysis. Also, two extra days (06/01/06 and 06/08/06) were added in the month of June, 2006 to maintain consistency in statistical analysis, because post-

processing revealed that the two dates 6/11/06 and 06/24/06 collectively yielded less than half as many vocalizations as did each of the months of May, July, and August.

An extra day (04/09/07) was also chosen for the month of April, 2007, counter to the sampling strategy for the child's developmental period of 15-24 months, because the day originally chosen (4/27/07) heavily features the child's newborn sibling. We study these two days separately to control for any anomalous patterns that may be due to the extraordinary events on April 27.

3.2.3.4 Speaker ID Filtering Analysis

As described in Section 3.2.3.1, we applied speaker ID metadata to filter out irrelevant segments from the annotation dataset in order to minimize overall annotation time. In the context of this work, irrelevant segments are those that do not contain child vocalizations. Because the focus of annotation was to capture as many child vocalizations as possible, we wished to retain not only segments labeled as the child, but also segments labeled as an adult or *other* speaker if they contain a child vocalization. The speaker ID algorithm assigns a confidence value to each segment by subdividing the segment into windows, determining a speaker label for each window, and using a voting mechanism across these windows to see which speaker is most prevalent in the segment. This means that a segment labeled as an adult or *other* can indeed contain child vocalizations, especially if the confidence is fairly low. To minimize the number of such child vocalizations that are lost by filtering out adult/other segments, we filtered them out only if their speaker ID confidence value was greater than a certain *confidence threshold* chosen for that speaker, as described in more detail below.

In implementing our filtering operation, the problem was to determine the optimal set of confidence thresholds across speakers that maximizes both the effectiveness and accuracy of this filtering process. We conducted a parameteric analysis to determine these optimal confidence thresholds for each of the non-child speakers. We denote the resulting set of four optimal confidence thresholds as the optimal *confidence threshold configuration*. Our criteria in determining optimality in this filtering problem were as follows: if the thresholds are too high, very few adult/other segments would be filtered out, offering no

time savings and making the filtering operation useless. This was the more conservative option, as leaving in adult segments would simply give annotators more segments to listen to. Because the clip for an irrelevant segment contained no child vocalizations, annotators would simply move on to the next clip with no harm done except for the extra time required to listen to it. However, if the thresholds are too low, too many adult/other segments containing child vocalizations would be filtered out, leaving out useful vocalization instances from our analysis.

Thus, an optimal confidence threshold configuration is one that minimizes the number of child vocalizations filtered out, while at the same time providing a useful reduction in the number of segments submitted for annotation by filtering out irrelevant segments. In the context of this filtering problem, we define the following terms:

- **True Positives (TP)** – segments containing child vocalizations that the filtering process accepts into the filtered annotation dataset

- **False Positives (FP)** – segments that do not contain child vocalizations, but are still accepted into the filtered annotation dataset

- **True Negatives (TN)** – segments not containing child vocalizations that are correctly filtered out of the annotation dataset

- **False Negatives (FN)** – segments that contain child vocalizations, but are filtered out of the annotation dataset

Restating the problem in these terms, the goal is to minimize the number of false negatives, while still providing an effective filter by filtering out false positives.

Minimizing false negatives is equivalently described as maximizing the *sensitivity* of the filtering configuration, where sensitivity is computed as the ratio of segments correctly accepted as containing child vocalizations (TP) to all the segments that indeed contain vocalizations (TP+FN):

$$sensitivity = \frac{TP}{(TP + FN)}$$

Analogously, minimizing false positives can be equivalently described as maximizing the *specificity* of the filtering configuration. Specificity is defined as the ratio of irrelevant segments correctly filtered out (TN) to all irrelevant segments (FP+TN):

$$specificity = \frac{TN}{(FP + TN)}$$

For any given confidence threshold configuration, we determined the optimal confidence threshold configuration as follows. First, we chose a set of already annotated speech segments for the analysis. For each speech segment in this set, we retrieved its assigned speaker ID label and confidence value. If the speaker ID was labeled as the child, then we immediately accepted this segment into the annotation dataset. Otherwise, if the confidence value was greater than the threshold configured for that speaker, then this segment was rejected from annotation: with a non-child speaker ID label that was hereby considered accurate, the segment was not likely to contain child vocalizations. Otherwise, the segment was accepted into the annotation dataset, because there was a chance that child vocalizations may be present.

We then queried annotated answers to clip-level question #1 ("Does this clip include child vocalizations?") for this set of segments, to serve as the perceptual ground truth for the evaluation. An answer of "yes" for a segment marked "accepted" were tallied as TPs, while an answer of "no" added to the count of FPs. Analogously, "no" answers for segments marked "rejected" were tallied as TNs, while "yes" answers added to the number of FNs.

We repeated this process for a variety of confidence threshold configurations and computed sensitivity and specificity for each configuration, using its corresponding TP, FP, TN, and FN totals. We plotted an ROC curve (1-specificity vs. sensitivity), as well as filtering ratio as a function of sensitivity, to evaluate the tradeoff between accuracy and effectiveness of the filtering process. Using these graphs as a guide, we chose a desired range of filtering ratios. We then ranked all threshold configurations falling within this

range by sorting along three criteria: first, in order of decreasing filtering ratio; second, in order of decreasing sensitivity; and third, in order of decreasing specificity. Finally, we chose the first threshold configuration in the ranking as the optimal one for the specified range of filtering ratios.

Six days' worth of speech segments were already fully annotated when speaker ID metadata became available: 5/16/06, 5/21/06, 6/11/06, 6/24/06, 7/2/06, and 7/10/06. In our parametric analysis method described above, we used this set of segments to indicate which segments actually contain child vocalizations and which do not. We assembled a complete set of confidence threshold permutations, with confidence thresholds ranging from 0 to 1 at intervals of 0.05, for each speaker label besides the child. There were four speaker labels under consideration: *father*, *mother*, *nanny*, and *other*. With 20 possible confidence thresholds for each of four speaker labels, we analyzed a total of 160,000 different threshold configuration permutations.

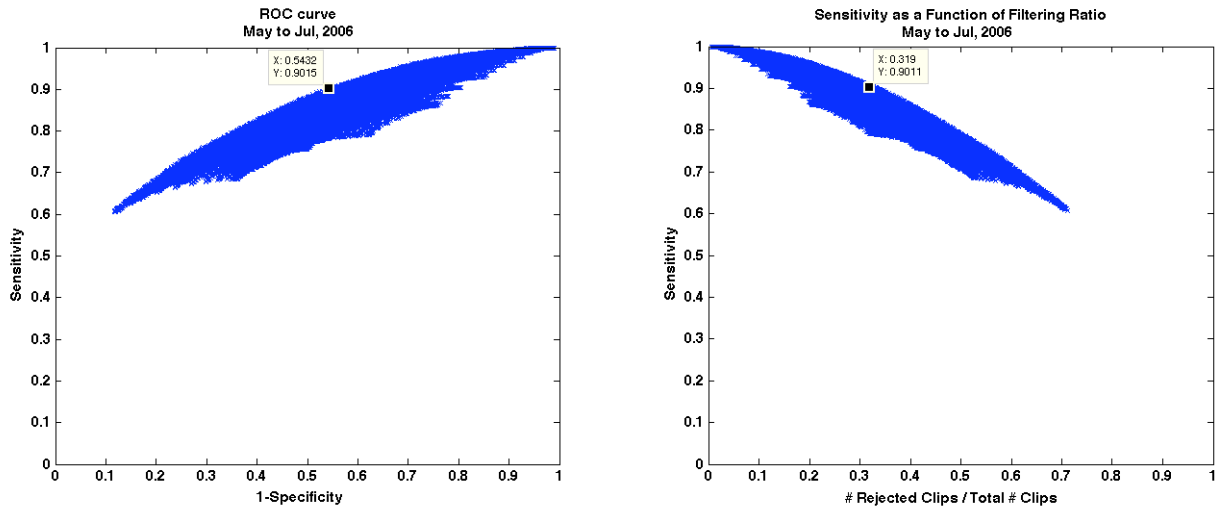


Figure 3-9. Evaluation of tradeoffs between Sensitivity, Specificity, and Filtering Ratio for different confidence threshold configurations: (a) ROC Curve (b) Sensitivity as a Function of Filtering Ratio. Each individual data point in these graphs is a different confidence threshold configuration. The goal is to find an optimal confidence threshold configuration for filtering out adult speech from the annotation dataset using Speaker ID labels and confidence thresholds. A confidence threshold configuration is a set of four confidences, one for each non-child speaker, that each serve as cutoff thresholds for their corresponding speaker. Segments labeled as that speaker are rejected if Speaker ID assigned a confidence greater than the confidence threshold for that non-child speaker.

Confidence Threshold Configuration				# Segments After Filtering		Filtering Accuracy						
Father	Nanny	Mother	Other	Accepted	Rejected	TP	FP	TN	FN	$\frac{TP}{(TP+FN)}$	$\frac{FP}{(FP+TN)}$	Filter Ratio
0.60	0.75	0.45	0.55	5500	2356	2496	3004	2121	235	0.914	0.586	0.30
0.55	0.75	0.35	0.65	5107	2749	2412	2695	2430	319	0.883	0.525	0.35
0.70	0.50	0.40	0.45	4714	3142	2327	2387	2738	404	0.852	0.466	0.40
0.35	0.70	0.35	0.50	4321	3535	2239	2082	3043	492	0.820	0.406	0.45
0.55	0.45	0.25	0.50	3929	3927	2127	1802	3323	604	0.779	0.351	0.50
0.35	0.40	0.45	0.40	3536	4320	2015	1521	3604	716	0.738	0.297	0.55

Table 3-1 Optimal Confidence Threshold Configurations. A confidence threshold configuration is a set of four confidences assigned by Miller's speaker ID algorithm – one for each non-child speaker. Segments assigned a confidence higher than the confidence threshold for that speaker are rejected from the annotation dataset as irrelevant, because they are deemed likely to be purely that speaker, and therefore would not contain any child vocalizations. The accuracy and effectiveness of this assumption for a given confidence threshold configuration is evaluated in terms of sensitivity and specificity, using as ground truth the set of segments that had already been annotated at the time. A different optimal confidence threshold configuration is derived depending on the desired filtering ratio. Highlighted in yellow is the configuration chosen for filtering the rest of the dataset.

Figure 3-9 shows the resulting ROC and Filtering Ratio graphs. Each data point in these graphs corresponds to a particular threshold configuration permutation. Determining the optimal tradeoff primarily depends on how much sensitivity one is willing to give up to gain an extra 5 or 10% of dataset reduction, and therefore annotation time savings.

An optimal configuration, chosen from a set of viable options, was determined as shown in Table 3-1, with a filtering ratio of 0.3. This was a conservative choice, made with the intention of keeping the number of false negatives at a minimum while still obtaining some useful reduction in annotation volume. Table 3-1 also lists several other possible options that were considered, which exchange some sensitivity for a larger filtering ratio. The parametric analysis described above showed each of these options to be the clearly optimal ones for their respective filtering ratios.

Using the chosen threshold configuration, the speaker ID filtering process was applied to the remainder of the input annotation dataset, which was still awaiting annotation. Table 3-2 shows the resulting reduction in annotation volume for each day's worth of segments. The dates that were annotated after filtering realized the benefits of the reduction. These dates are highlighted in Table 3-2. The filtering ratio is lower on average for the rest of the dates, than for the dates that were used to compute the threshold

configuration, and ranges from as low as 0.175 to a high of 0.311. Overall, the speaker ID filtering process reduced the amount of segments to be heard by annotators by 22%.

Date	Total # Segments	Accepts	Rejects	Filtering Ratio
05/16/2006	1807	1336	471	0.260
05/21/2006	1064	750	314	0.295
06/11/2006	1361	919	442	0.325
06/24/2006	791	554	237	0.300
07/02/2006	1220	845	376	0.308
07/10/2006	1619	1103	516	0.319
08/07/2006	1323	1003	319	0.241
08/22/2006	2014	1607	407	0.202
09/06/2006	1118	807	311	0.278
09/17/2006	2130	1468	662	0.311
10/05/2006	1982	1622	360	0.182
10/19/2006	1884	1554	330	0.175
01/04/2007	1755	1423	332	0.189
04/09/2007	3227	2471	756	0.234
04/27/2007	2689	2121	568	0.211
07/06/2007	2180	1869	311	0.143

Table 3-2. Applied Speaker ID Filtering Results. Highlighted in blue are the days for which the chosen confidence threshold configuration was applied to filter out irrelevant segments from the annotation dataset. Filtering out 100 segments saves roughly 2 hours of annotation time. Overall, filtering by speaker ID reduced the amount of segments to be heard by annotators by 22%, for a total time-savings of roughly 87 hours.

In addition to the filtering ratio, the raw number of rejected segments also tells us how much annotation time speaker ID filtering saved, regardless of the total number of segments in a given day. In practice, each set of 100 segments takes roughly an hour to annotate. With each segment being annotated by two annotators, filtering out 100 segments saves 2 hours of annotation time. Thus, filtering by speaker ID saved as many as 14-16 hours of annotation time for a given day (e.g., rejecting 756 segments in the 04/09/07 data), and provided an average time-savings of 9 hours per annotated day. In total, speaker ID filtering reduced annotation time by roughly 87 hours.

3.3 Annotation Process

Annotation took place from January through May 2010. Seven undergraduate students were recruited through MIT's Undergraduate Research Opportunities Program to serve as annotators. Their demographics are listed in Table 3-3, with numerical identifiers assigned for confidentiality.

ID	Gender	Class Year	School	Major	Started
1	Female	Freshman	MIT	Math	1/10
2	Female	Freshman	MIT	Biology	1/10
3	Male	Freshman	MIT	EECS	2/10
4	Male	Sophomore	MIT	Math/EECS	1/10
5	Female	Sophomore	MIT	EECS	1/10
6	Female	Junior	MIT	Brain & Cognitive Science	1/10
7	Female	Senior	Wellesley	Neuroscience	2/10

Table 3-3. Annotator Demographics.

All annotators underwent a period of adjustment and training. The first two weeks in January involved testing out the annotation interface and supporting infrastructure. During this time, annotators were given a test dataset and a variety of event types to annotate and questions to answer. By completing these assignments, annotators gained experience in interface usage. In the second half of January, a preliminary version of the actual input configuration (trackmap and questionnaire) used in this work was deployed to the annotators. During this preliminary trial, the paradigm for distribution of assignments across annotators was also established: each assignment, consisting of 100 segments[7], was assigned to two annotators so that inter-annotator agreement could later be measured. After roughly three weeks of pilot testing and revisions of the questionnaire, we deployed the final version of the input configuration described in Section 3.2 in early February.

At this point, evaluation of inter-annotator agreement and interviews with the annotators revealed the challenges involved in achieving agreement on annotators' subjective perceptions of a child's affective state. Across the board, annotators expressed difficulty in judging how extreme (or not) a child's energy and mood are relative to the

[7] Each day's worth of segments was subdivided into groups of 100 to create these assignments. The last assignment for a day contained less than 100 segments -- specifically, the remainder after this subdivision.

child's overall range. Also, definitions were needed to clarify the distinctions between **crying**, **laughing**, **babble**, and **other emoting** for answers to annotation-level question #1 ("Which of the following best represents the nature of the vocalization?").

To address these problems, two measures were taken to create a formal training process for the annotators. First, a one-page instruction sheet was provided to the annotators as a quick reference to definitions, rules of thumb, and example social situation cues. A copy of this instruction sheet is included as Appendix C. The definitions established for the choices in annotation-level question #1 are as follows:

> **Crying** - in a child, it is an inarticulate, often prolonged expression of a negative state, and can range from soft weeping to screaming, depending on the energy of expression. It can also begin with whining or other fussy vocalizations.
>
> **Laughing** - expressing certain positive emotions, especially amusement or delight, by a series of spontaneous, usually unarticulated sounds, such as heehee, hehe, haha.
>
> **Bodily** - a vocalization that is produced by reflexive bodily functions and therefore carries no emotional content, such as a sneeze, cough, or burp.
>
> **Babble** - a vocalization where the child is attempting to express speech, with or without emotion. It must include at least two clearly articulated consonant-vowels, i.e. at least two clear syllables.
>
> **Speech** - the child is clearly speaking in articulated, recognizable words.
>
> **Other Emoting** - any other vocalization uttered by the child

As can be expected within the continuous spectrum of a child's emotional expressions, annotators observed many borderline cases where they found it hard to tell whether the child was crying or simply emoting, and similarly, whether the child was laughing or emoting. To address this ambiguity, annotators were instructed to choose **other emoting** to characterize these borderline cases.

The second measure was a training process created to familiarize the annotator with the child's overall temperament and range of expression. Three sets of clips were put together, one for each of annotation-level questions #1 (Nature), #2 (Energy), and #3 (Mood). Each set presented a range of canonical examples for each of the options in the corresponding question. The example clips were chosen to represent the full range of the child's expressive tendencies, as well as a good spectrum of possibilities within each rating option. Each clip was labeled with the question and the suggested answer for that clip.

An interface was implemented for browsing the labeled clips in each set (Appendix A, Figure 4), creating three "guides" for acquainting annotators with the child's temperament[8]. Annotators were asked to browse through these guides from beginning to end. The guides were also added to the Q&A interface (Appendix A, Figure 3), each under its corresponding question. This was done to enable the annotator to refer back to the guides while annotating with just a press of a button, to help disambiguate difficult cases.

All annotators finished this training process by February 22. Speaker ID metadata became available in mid-February, and after the analysis in Section 3.2.3.4 was done, input datasets filtered using speaker ID were introduced for annotation the first week of March.

3.4 Agreement Analysis

In this section, we evaluate the inter-annotator agreement for the four questions in the questionnaire that involve subjective judgment on the part of the annotator. Originally described in Section 3.2.2, these questions are restated in Table 3-4 and mapped to a one-word label that will henceforth be used to refer to these questions, for ease of exposition.

Agreements are computed as follows. First, all individual child vocalization annotations and the question responses for each annotation are queried and assembled into **ResultRecord** objects. A **ResultRecord** associates a time interval and question label with the choices that annotators selected for this time interval and question. Using a Hashtable, the **ResultRecord** maps each choice option to a Vector of annotator usernames who selected that choice. Initially, a distinct **ResultRecord** object is created for each individual child vocalization annotation in the database. In other words, the Hashtable of choices in each initial **ResultRecord** contains only a single key – the option chosen by the annotator who made this annotation – and a Vector of size 1 that contains only the username of that annotator.

[8] Thanks to Jason Hoch, one of the annotators, for his contributions to the implementation of the annotator training guide interface.

Label	Question	Options
Social	Is there any activity in this clip that might suggest a social situation involving the child and a caretaker? (See instruction sheet for examples.)	Yes, No
Nature	Which of the following best represents the nature of the vocalization?	Crying, Laughing, Bodily, Babble, Speech, Other Emoting
Energy	Please rate the energy of this vocalization. If it varies within the vocalization, rate its maximum energy: (1 = Lowest energy, 5 = Highest energy)	1, 2, 3, 4, 5
Mood	Please rate the child's mood on the following scale: (1 = Most negative, 5 = Most positive)	1, 2, 3, 4, 5

Table 3-4 Subjective Questions Evaluated in Agreement Analysis.

Then, individual ResultRecords whose time intervals intersect with each other are aggregated into a single ResultRecord. Non-intersecting portions are split off (and corresponding Hashtable contents are copied over) into separate ResultRecord objects. Splitting is necessary to handle the case where one annotator may have indicated multiple brief vocalization intervals, while another annotator may have simply annotated one long vocalization, for the same period of time. In the former case, each of the vocalizations may have been given a different set of question responses, while in the latter case, there is only one set of responses for the entire period. Thus, one portion of this time period may hold an agreement, while another portion may hold a disagreement. With splitting, each of these portions would be aggregated into separate ResultRecord objects, and would duly capture the agreement as well as the disagreement.

Finally, the number of agreements is tallied within this set of aggregated ResultRecords. For a given question, an agreement is defined to be a vocalization interval for which two or more annotators made the same choice. This was expressed programmatically by checking whether a ResultRecord's Hashtable contains at least one Vector value of size ≥ 2. Because this Vector contains the usernames of the annotators who made the same choice, size ≥ 2 means that at least two annotators agreed with each other.

For the scale type questions, namely Energy and Mood, we consider agreement across four different rating metrics: the original 5-point scale $\{1, 2, 3, 4, 5\}^{5pt}$, two versions of a

collapsed 3-point scale with options denoted as {1, 2, 3}^{3pt}, and a 9-point scale {1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5}^{9pt} resulting from accepting disagreements by 1 point as agreements and averaging them to obtain the new agreed value. The two versions of the 3-point scale differ according to how the intermediate values 2 and 4 in the original 5-point scale are categorized in the collapsed 3-point system, with specific mappings as follows:

- 3-point Scale Version A: {1,2}^{5pt} → {1}^{3pt}, {3}^{5pt} → {2}^{3pt}, and {4,5}^{5pt} → {3}^{3pt}
- 3-point Scale Version B: {1}^{5pt} → {1}^{3pt}, {2,3,4}^{5pt} → {2}^{3pt}, and {5}^{5pt} → {3}^{3pt}.

Table 3-5 summarizes the inter-annotator agreement for each question and scale type variant. Agreement is measured according to two metrics: probability of agreement **P(Agreement)**, and **Cohen's Kappa**, which adjusts for any role of random chance in contributing to agreement (Cohen, 1960). Cohen's Kappa is calculated using the formula:

$$\kappa = \frac{P(Error) - P(Agreement)}{1 - P(Error)}$$

where *P(Error)* is the probability of two annotators agreeing by chance, as if they were effectively flipping coins to make their ratings. Cohen's Kappa therefore provides an agreement metric that tells us not only how much annotators agreed with each other, but also the extent to which that agreement meaningfully represents a correlation between annotator ratings and the characteristics of the child vocalization events being annotated. We interpret the value computed for Kappa according to the following commonly used rating system (Altman, 1991):

Poor agreement = Less than 0.20
Fair agreement = 0.20 to 0.40
Moderate agreement = 0.40 to 0.60
Good agreement = 0.60 to 0.80
Very good agreement = 0.80 to 1.00

Label	Total # Child Vocalizations	# Agreed	P(Agreement)	P(Error)	Cohen's Kappa (κ)
Social	11518	9518	0.826	0.612	0.552
Nature	11470	7972	0.695	0.391	0.499
Energy (5 pt scale)	11543	5537	0.480	0.253	0.304
Energy (3 pt scale A)	11543	6996	0.576	0.317	0.422
Energy (3 pt scale B)	11543	9470	0.820	0.712	0.376
Energy (9 pt scale)	11543	10543	0.913	0.142	0.899
Mood (5 pt scale)	11544	5838	0.506	0.273	0.321
Mood (3 pt scale A)	11544	6856	0.593	0.306	0.413
Mood (3 pt scale B)	11544	10298	0.892	0.840	0.324
Mood (9 pt scale)	11544	11000	0.953	0.177	0.943

Table 3-5. Agreement Calculations. Very good agreement is highlighted in yellow (9 point scale for both Energy and Mood). Moderate agreement is highlighted in green.

As can be seen in Table 3-5, the Social and Nature questions achieve moderate agreement, with Kappa values of 0.522 and 0.499, respectively. Among the four scale type variants for the Energy and Mood questions, the 9-point scale seems to be by far the most robust option for further analysis: both questions achieved very good agreement, with Kappa values of 0.899 and 0.943, respectively. The original 5-point scale achieved fair agreement, with Kappa values of 0.304 and 0.321, respectively. Of the two 3-point scales considered, version A was superior to version B for both Energy and Mood: version A achieved moderate agreement with Kappa values of 0.422 and 0.413, respectively, while version B had only fair agreement with Kappa values of 0.376 and 0.324, respectively.

Agreement percentages per individual annotator for each question are included in Appendix D.

Chapter 4

Analysis Methodology

In our analysis methodology, we seek to create a model that simulates human perception of a child's emotional state, given the child's vocalizations and the adult speech surrounding the vocalization. Acoustic features of the child's vocalizations and surrounding adult speech serve as the independent (input) variables, and the <Mood, Energy> ratings made by annotators serve as the dependent (response) variables. In the process, we investigate the degree to which surrounding adult speech alone can indicate a child's perceived emotional state, how such correlations may differ for specific child-caretaker dyads, and the nature of any longitudinal trends in these dyadic relationships as the child develops from 9 to 24 months of age. We apply these research goals to different subsets of the data to explore social context as given by annotator responses to the Social question, as well as the nature of the vocalization given by answers to the Nature question. Henceforth, we define the combination of the Social and Nature answers for each vocalization to be the vocalization's *socio-behavioral context*.

This chapter describes our implementation of this analytical methodology, starting with the raw annotation data collected as described in Chapter 3. The analysis process involves three phases. First, a data processing phase takes the collected annotation data, creates indices for agreed Q&A answers, and generates four distinct sets of raw audio fragments from the Speechome data corpus, using the time interval annotations for child vocalizations and overlapping noise/adult speech. These four sets of audio fragments are

(1) The child vocalizations themselves, pruned of any overlapping noise/adult speech

(2) Adult speech occurring within a 30 second window ***before*** each child vocalization,

(3) Adult speech occurring with a 30 second window ***after*** each child vocalization, and

(4) The concatenation of (2) and (3) into adult speech ***surrounding*** each child vocalization. Our implementation of this data processing pipeline is described in Section 4.1.

The second phase of analysis extracts a set of 82 acoustic vocal features from the raw audio fragments. Section 4.2 describes each of these features and the functionality that we developed to implement the feature extraction process.

Finally, we apply Partial Least Squares (PLS) Regression to build perceptual models of child emotion using these acoustic features. In order to address the research questions above, we built multiple different models, each using a different longitudinal or socio-behavioral subset of the data. Section 4.3 describes our rationale for choosing PLS Regression, the experimental designs that implement this regression methodology, and our metric for evaluating the regression models.

4.1 Data Processing

The data processing pipeline can be summarized as follows. First, the annotated child vocalization time intervals are pruned to remove any periods of overlapping noise or adult speech. Next, WAV audio files are generated from the Speechome corpus for each pruned child vocalization. The audio fragments for surrounding adult speech are then also derived, and WAV files generated, using the pruned child vocalization time intervals as templates. In parallel, indices of agreed answers are created for each of the four questions in Table 3.4. The start and end times of the pruned vocalization intervals serve to uniquely index each generated audio fragment and agreed Q&A answer according to its corresponding child vocalization. Finally, the WAV audio files are submitted to the feature extraction process (see Section 4.2).

4.1.1 Pruning

Prior to pruning, we compute all time interval intersections across annotators in the set of all collected annotations. This is done separately for child vocalization annotations and overlapping noise/adult speech, in a process of aggregation and splitting of ResultRecords that is very similar to, and in fact a direct extension of, the implementation described in Section 3.4. Instead of tallying Q&A agreements, here we aggregate only the time interval annotations within each event type. Since each clip could have multiple vocalization and noise annotations, the agreed time intervals occurring during each clip are organized in vector sets that are mapped, using a Hashtable, to the corresponding **clip_id**.

Pruning is then done by iterating through the set of agreed noise intervals, searching for any intersections within the set of child vocalization intervals, and pruning these intersections from the vocalization intervals. Indexing both sets of agreed time intervals – child vocalizations and overlapping noise/adult speech – by **clip_id** enabled us to implement an efficient pruning algorithm by reducing this search space to just the small set of aggregated and split annotations occurring within a single clip.

For each intersection found between a noise interval and a vocalization interval, there are four distinct cases to consider, each one involving a different pruning operation:

1. **The intersection is at the beginning of the vocalization interval.** We simply prune the noise off by changing the start time of the vocalization to the end time of the noise interval.

2. **The intersection is at the end of the vocalization interval.** Conversely to case 1, pruning is done by changing the end time of the vocalization to the start time of the noise interval.

3. **The intersection is somewhere in the middle of the vocalization interval**. In this case, the noise splits the vocalization interval into two fragments. We adjust either the start or end time (but not both) in the original vocalization interval to

reflect the first vocalization fragment, and construct an additional ResultRecord to hold the second vocalization fragment created by this split.

4. **There is a complete overlap between the noise interval and vocalization interval.** The vocalization is removed completely from the analysis dataset.

After this pruning operation, the resulting adjusted and/or split vocalization intervals are checked whether they are "zero-length", based on a configured threshold, and if so, they are removed from futher analysis. In this work, "zero-length" is defined as being shorter than 100ms, so that vocalization fragments are thrown out if they are too short to be perceptually meaningful to a human listener. Otherwise, they are added to the master set of pruned vocalizations to be analyzed from this point forward.

4.1.2 Generating WAV files

In preparation for feature extraction (Section 4.2), we extract the raw audio from the Speechome corpus, and generate a WAV audio file, for each specific time interval that we have computed in this master set of pruned child vocalizations. This extraction and conversion process was implemented by applying input-output functions already existing and available for use in the Speechome software library to extract specific time intervals of audio from the Speechome corpus, as well as WAV encoding functions in the JavaZoom AVS Audio Converter API[9].

4.1.3 Surrounding Adult Speech

As part of this work, we are studying what adult speech surrounding a child vocalization can reflect about the child's perceived emotional state. To this end, we implemented a module called GenWavsAdult that computes the time intervals for adult speech surrounding each child vocalization and generates WAV files for these intervals.

[9] http://www.javazoom.net/index.shtml

Specifically, **GenWavsAdult** returns all adult speech occurring within a user-specified window before and after each child vocalization in the form of WAV files, and an index mapping the adult speech intervals to child vocalizations. The output can include all adult speech, or focus in on only the speech of a particular adult, to enable a dyadic analysis between the child and a specific adult.

GenWavsAdult begins its derivation of adult speech intervals surrounding each child vocalization by retrieving all the pruned child vocalization time intervals within a particular date range, specified by input parameters <dateRangeStart> and <dateRangeEnd>. For each child vocalization, specific time intervals T1 and T2 are computed to cover a user-specified time window before the child vocalization and after, respectively. In this work, both window lengths have been set to 30 sec, as a time interval that reasonably captures the child's attention span. It seems reasonable to expect that any emotional transitions relevant to a particular vocalization would evolve within the time frame of the child's attention span before and after that vocalization.

Given a computed window T (this is done separately for T = T1 and T = T2), the task now is to find all speech segments that overlap with T. Because **GenWavsAdult** can be run on the entire corpus of 1.6 million segments, and was initially implemented with the intention to do so, an **IntervalTree** – an augmented binary search tree with red-black balancing (Cormen et al., 2001) – was implemented from scratch and applied to improve the efficiency of the search.

The set of all speech segments occurring between <dateRangeStart> and <dateRangeEnd> in the Speechome database is first organized into an **IntervalTree** structure S. A separate **IntervalTree** C is also built out of all the pruned child vocalization intervals. Then, the window T is submitted to the **IntervalTree** search function on tree S, which returns a set V of all segments that overlap with T. Finally, for each segment in V, the exact overlapping time interval I is computed, and using the **IntervalTree** C of child vocalizations, any child vocalizations occurring during interval I are pruned out.

The end result of this process is a set of audio fragments that collectively represent all audio of human speech within window T that is not produced by the child. This set of adult

speech fragments in window T is saved in a Hashtable that maps it to the child vocalization from which T was computed. Thus, for T = T2 with a window length of 30 sec, a given child vocalization will be mapped to the set of adult speech fragments that occur during the 30 seconds before this vocalization.

Finally, the audio corresponding to each adult speech fragment is retrieved from the Speechome corpus and written to a WAV file, as in Section 4.1.2. Also, an index file is written, which contains all the mappings between child vocalizations and their corresponding adult speech fragments.

4.1.4 Generating Agreement Indexes

For each of the four subjective questions in Table 3-4 (Social, Nature, Mood, and Energy), including the multiple scale variants for the Mood and Energy questions, we create an agreement index that maps all agreed answers to the pruned vocalization intervals to which they correspond. In the PLS Regression analysis (see Section 4.3), we use these indices to compute subsets of vocalizations that correspond to agreed answers for each question, and to coordinate their prosodic features with their matching answers for the regression. Agreements among answers for each pruned child vocalization are computed according to the methods described in Section 3.4.

4.2 Feature Extraction

The feature extraction process takes the generated WAV files and computes a set of 82 acoustic and prosodic features for each child vocalization and its adjacent windows of adult speech. The features are computed via Praat, a widely used toolkit for speech analysis and synthesis (Paul Boersma & Weenink, 2010). The features extracted for analysis are summarized in Table 4-1 and described in greater detail in Section 4.2.1.

Attribute	Metrics	Units	Interpolation
Intensity	Min, Mean, Max, Stdev	dB	Parabolic
Pitch	Min, Mean, Max, Stdev	Hertz	Parabolic
	Mean absolute slope	Semitones	n/a
Stylized Pitch Contour	# stylized pitch points	Quantity	n/a
Fast Fourier Transform (FFT)	Centroid, Stdev	Hertz	n/a
	Skewness, Kurtosis	Unitless	
Long-Term Average Spectrum (LTAS)	Min, Mean, Max, Stdev	dB	None
	Frequency of Min, Frequency of Max	Hertz	None
Harmonics-to-Noise Ratio (HNR)	Min, Mean, Max, Stdev	dB	Parabolic
Harmonics-to-Noise Ratio (HNR)	Time of Min, Time of Max	Seconds	Parabolic
Mel-Frequency Cepstral Coefficients (MFCCs), #1 – 16	Mean, Stdev (separately for each MFCC)	Mel	n/a
Formants, #1 – 5	Min, Mean, Max, Stdev	Hertz	Parabolic

Table 4-1. Acoustic Features Extracted for Analysis

4.2.1 Features

This section describes the specific features extracted for analysis within the context of Praat syntax. As shown in Table 4-1, the set of 82 features is constructed using eight attributes of the audio signal. Two of the attributes – MFCCs and Formants – characterize an audio signal across multiple coefficients or dimensions. To obtain our features, we read each input WAV file into a Praat **Sound** object, and then compute multiple attribute-specific metrics accessible through the Praat API to quantify each of these attributes or attribute dimensions. Metrics include minimum, mean, maximum, standard deviation, and various others, as listed in Table 4-1.

4.2.1.1 Intensity

The Praat **Intensity** object provides query access to the intensity contour for a given utterance. Using the Praat API of available query functions, we compute four intensity-related features for our analysis, with parabolic interpolation: the minimum, mean, maximum, and standard deviation of the intensity contour.

The following syntax is used to create a Praat **Intensity** object from a **Sound** object, with parameters set to their default values as recommended by Praat (Paul Boersma & Weenink, 2010):

```
To Intensity... <min_pitch> <time_step> <subtract_mean?>
```

with `<min_pitch>` set to 100 Hz, `<time_step>` set to 0, and `<subtract_mean?>` set to "yes". The `<min_pitch>` parameter specifies the minimum periodicity frequency, which determines the smoothing window for removing pitch-related variations in intensity. The `<time_step>` parameter specifies the time step used in the resulting intensity contour; set to 0, as per the default, the time step is computed as one quarter of the effective window length: 0.8/*min_pitch.* Set to "yes", the `<subtract_mean?>` parameter normalizes the intensity contour by removing any DC offset that may have been introduced as an artifact of microphone recording.

4.2.1.2 Pitch

We compute five pitch-related features: minimum, mean, maximum, standard deviation, and mean absolute slope. Our motivation for including the latter in our set of features, in addition to the first four self-explanatory metrics, is that it tells us the general trend of the pitch contour – does the voice generally rise or fall in the child vocalization or adult response? Although mean absolute slope is a rather aggregate representation of the pitch contour, it is widely used for this purpose, particularly in the context of affect analysis (Breazeal, 2001; Chuang & Wu, 2004; Liscombe et al., 2005; Slaney & McRoberts, 1998; Ullakonoja, 2010). Section 4.2.1.3 describes an additional aggregate feature that we derive from the pitch contour to characterize its time-varying behavior from a more fine-grained perspective.

We generate the pitch contour by instantiating a Praat **`Pitch`** object from a **`Sound`** object, using the following syntax:

```
To Pitch... <time_step> <pitch_floor> <pitch_ceiling>
```

The time step and pitch floor are set to the default value of 0.01 sec and 75 Hertz, respectively. Together, pitch floor and pitch ceiling define the pitch range. The standard pitch range for adults is from 75 to 600 Hertz (Paul Boersma & Weenink, 2010; Petroni,

Malowany, Johnston, & Stevens, 1994; Ullakonoja, 2010). Our choice of pitch ceiling takes into account the kinds of extremes in high pitch that can occur when a young child cries or squeals in excitement, as well as the imitative rise and exaggerated range expansion in adult pitch that is characteristic of "motherese" (Dominey & Dodane, 2004). The pitch ceiling filters out any pitches above this value, if they occur.

To adequately capture a full range of affective extremes, we set the pitch ceiling to a generous 2000 Hz (Clement et al., 1996; Petroni, Malowany, Johnston, & Stevens, 1994; Zeskind, 2005). Such a large pitch range is inevitable in the study of child emotion, but it bears the brunt of the tradeoff between undersampling with a frame size that is too short relative to the pitch floor and losing the stationarity property with a frame size that is too long relative to the pitch ceiling. Using a pitch floor of 75 Hz in the context of high-pitched child vocalizations, this tradeoff is tipped in favor of nonstationarity, which could lead to smearing of F0 values at the high end of the spectrum.

4.2.1.3 Stylized Pitch Contour

To capture some meaningful information, in aggregate, about the time-varying pitch movements during a child vocalization or surrounding adult speech, we include a feature that is inspired by the pitch contour stylization approach of Hart et al (1990). Pitch contour stylization is a technique that reduces the curves of a pitch contour to a set of straight-line segments that approximate the large-scale pitch variations within an utterance. From the stylized representation, it is then possible to compute the number of major direction changes indicated by the varying slopes of the stylized segments. This can be done by counting the number of segments in the stylization, or equivalently, the number of *stylized pitch points*, or vertices separating the segments.

We stylize the pitch contour by calling Praat's **`PitchTier:Stylize...`** function with a frequency resolution of 2 semitones, and then compute the number of stylized pitch points using the query function **`PitchTier:Get number of points`**. Both functions are part of the API available in Praat's **`PitchTier`** object, which represents a time-stamped pitch contour. Please refer to the example Praat scripts in Appendix E for more details on the syntax used for this procedure.

4.2.1.4 Fast Fourier Transform

The Fast Fourier Transform (FFT) is an operation that takes the time-varying amplitudes of the audio signal and maps the spectrum of frequencies that occur, effectively converting the signal representation from the time domain to the frequency domain (McClellan et al., 1999). The FFT thus facilitates analysis of the spectral properties of frequencies occurring within the audio signal. In Praat, we obtain the FFT using the **`Sound:To Spectrum…`** function. We compute four metrics using the FFT, as described below:

- **Centroid**, or ***center of gravity*** – the average frequency in a spectrum.

- **Standard Deviation** – measures how much the frequencies in the spectrum deviate from the Centroid.

- **Skewness** – measures how much the shape of the spectrum below the Centroid is different from the shape above it.

- **Kurtosis** – which measures the extent to which the shape of the spectrum around the Centroid is different from a Gaussian shape.

Praat allows us to compute each of these metrics using two kinds of weights: (1) the absolute spectrum, where the absolute value of the complex spectrum |S(f)| is used directly as the weight for frequency f; and (2) the power spectrum, which raises |S(f)| to the power of 2. Thus, the FFT gives us a total of eight spectral features.

4.2.1.5 Long-Term Average Spectrum (LTAS)

Long-Term Average Spectrum (LTAS) is defined as the logarithmic power spectral density (PSD) as a function of frequency. This produces a histogram of frequencies that tells us which frequencies are most or least dominant in the signal, and the degree to which

each frequency plays a role. We compute LTAS from the FFT using Praat's **Spectrum:To Ltas…** function with a bandwidth of 125 Hz, as per the settings used in Kovacic and Boersma (2006) and Kovacic et al (2003). The features that we extract from the LTAS are the minimum, mean, maximum, and standard deviation of the log PSD, measured in decibels (dB), as well as the frequencies of the minimum and the maximum, in Hertz.

4.2.1.6 Harmonics-to-Noise Ratio (HNR)

Another parameter demonstrated to be meaningful in characterizing emotion in the human voice is Harmonics-to-Noise Ratio (HNR). Various studies have found correlations between emotional content in speech and HNR (Klasmeyer & Sendlmeier, 1995; Tato et al., 2002) and, using HNR as part of the feature space, have built classifiers for spoken affect with some success. Yumoto et al (1984) used HNR as a metric for the degree of hoarseness of an utterance. Rank and Pirker (1998) described HNR as representing the "breathiness" in the voice, and applied it as one of the parameters for synthesizing emotional speech.

We compute HNR as a function of time using built-in Praat functionality that performs an acoustic periodicity detection on the basis of a forward cross-correlation analysis (Paul Boersma & Weenink, 2010). The following syntax and parameters, applied to a Sound object, define this function call:

```
To Harmonicity (cc)… <frame duration>
                     <pitch_floor>
                     <silence_threshold>
                     <# periods per window>
```

The frame and window are two separate constructs in this function's internal cross-correlation algorithm, which is described in detail in (P. Boersma, 1993). The algorithm performs a short-term analysis – computation is done only on local sections, or *frames*, of the overall signal at a time. The frame duration determines the temporal length of each of these sections. *Window* refers to the Hanning window of the frame, to which the Fast Fourier Transform is applied as part of the algorithm. Frame duration is set to 30 ms and pitch floor to 75 Hz as in Section 4.2.1.2. The silence threshold is the signal amplitude below which values are considered to indicate silence. This parameter is set to the standard

default value of 0.1. For the last parameter, number of periods per window, we follow Praat's recommendation of 4.5 as the value that is best for speech (P. Boersma, 1993).

We query six aggregate features from the HNR time series for our analysis feature set: minimum, mean, maximum, standard deviation, as well as the times (relative to the timeline of the utterance) of the minimum and maximum values.

4.2.1.7 Mel-Frequency Cepstral Coefficients (MFCCs)

Mel-Frequency Cepstral Coefficients (MFCCs) form a model, using the mel-frequency scale, that takes into account the human auditory frequency response (Davis & Mermelstein, 1980; Ververidis & Kotropoulos, 2006). This has made MFCCs a popular choice in modeling the phonetic content of speech (Nwe et al., 2003), because they provide a better representation of the signal than do the simple frequency bands in the power spectrum (Ververidis & Kotropoulos, 2006). However, because MFCCs suppress the fundamental frequency (Jensen et al., 2009), which has been consistently found to be a significant correlate with spoken affect, there have been mixed results in applying MFCCs in the context of emotion recognition from speech (Beritelli et al., 2006; Nwe et al., 2003).

Nevertheless, MFCCs have been found to hold potential in classification of spoken affect (Kamaruddin & Wahab, 2008). We include the mean and standard deviation for each of 16 MFCCs in our feature vector, in part to evaluate their effectiveness in the context of child vocalizations, and also to harness their physiological realism in modeling speech from the perspective of human perception. We suspect that MFCCs would add a useful dimension that would complement the set of other attributes (Intensity, Pitch, FFT, LTAS, and Formants) in our feature space. It is human perception, after all, that is the gold standard for emotion recognition.

4.2.1.8 Formants

The vocal tract carries a set of important resonances called formants. Particular formant frequencies manifest themselves as peaks in the frequency spectrum of an utterance (Sundberg, 1977) and appear as groups of high-intensity harmonics in a

spectrogram (Smith, 2007). The formant frequencies are determined by the shape of the vocal tract as the lips, tongue, pharynx, and jaw move in the process of vocal expression (Ali et al., 2006; Sundberg, 1977).

In turn, the shape of the vocal tract is influenced by a person's emotional state (Oudeyer, 2003; Ververidis & Kotropoulos, 2006), e.g. as constrictions occur in the throat due to stress. For this reason, formants have been consistently included in feature vectors for synthesis and recognition of emotional speech (Ali et al., 2006; Burkhardt & Sendlmeier, 2000; Martland et al., 1996; Murray & Arnott, 1993; Schroder, 2001; Ververidis & Kotropoulos, 2006). The research done in this area has found formants to be useful acoustic correlates of emotion in speech. Other works have even applied formants to the study of emotive meaning in dog barks and other animal vocalizations (Clemins & Johnson, 2006; Molnar et al., 2008; Soltis, 2009; Soltis et al., 2005).

We use Praat's **Sound:To Formant (burg)…** function to compute the first 5 formants for each child vocalization and surrounding adult utterance. **To Formant (burg)…** is the variation of Praat's formant implementation that is recommended for general use. Internally (Paul Boersma & Weenink, 2010), it computes LPC coefficients with a Gaussian-like window, using the Burg algorithm, as given by Childers (1978).

Input parameters to this function are ***time step*** (seconds), ***maximum number of formants***, ***maximum formant*** (Hertz), ***window length*** (seconds), and ***pre-emphasis starting point*** (Hertz). Standard default values are used for all input parameters, except one: the ***maximum formant*** is set to 8000 Hertz, as recommended by Praat for a young child[10]. The time step is configured as 0.0 such that Praat sets it internally to 25% of the window length. Maximum number of formants is set to 5, because we are computing the first 5 formants. Window length is set to 25 ms, for which Praat uses a Gaussian-like window that is effectively 50ms long. The final pre-emphasis starting point parameter is left at the recommended 50 Hertz; pre-emphasis is described as a smoothing operation that optimizes the spectrum for formant analysis (Paul Boersma & Weenink, 2010).

Table 4-2 lists the main physiological correlates for each of the five major formants. For our feature set, we query the minimum, mean, maximum, and standard deviation of

[10] http://fonsg3.hum.uva.nl/praat/manual/Sound__To_Formant__burg___.html

each of the formants per utterance, with parabolic interpolation. This yields a total of 20 formant-based features.

Formant #	Main Physiological Correlates
1	Amount of opening in the jaw (Sundberg, 1977); Any constriction or expansion in pharynx and front half of the oral part of the vocal tract (Ali et al., 2006), including distance of the tongue from the roof of the mouth[11]
2	Shape of the body of the tongue (Sundberg, 1977); The frontness or backness of the highest part of the tongue during the production of the vowel + degree of lip rounding[4] (Ali et al., 2006); Constriction/expansion of the back vs. front of the tongue (Ali et al., 2006)
3	The position of the tip of the tongue (Sundberg, 1977); phonemic quality of specific speech sounds (Ali et al., 2006)
4, 5	Voice quality by various internal organs of the vocal tract (Ali et al., 2006; Sundberg, 1977)

Table 4-2. Physiological Correlates for Formants 1-5

4.2.2 Automated Feature Extraction Process

A set of Praat scripts is generated programmatically and then run using the command-line interface invocation of Praat. For child features, the set of Praat scripts is built using the list of pruned vocalizations computed in Section 4.1.1, one script for each vocalization. For adult features, the index of mappings between vocalizations and surrounding adult speech fragments given by GenWavsAdult (Section 4.1.3) is used instead. Example Praat scripts for a child vocalization as well as surrounding adult speech are included in Appendix E.

For the surrounding adult speech, the resulting Praat script first concatenates the fragments corresponding to each child vocalization, with periods of silence filling in any breaks due to pruned overlapping child vocalizations. The features described in Section 4.2.1 are then computed for the concatenated audio sequence. This process is identical for child vocalizations, except that each vocalization WAV file is submitted directly to Praat for feature extraction, with no need for concatenation of fragments.

[11] http://cslu.cse.ogi.edu/tutordemos/SpectrogramReading/ipa/formants.html

4.3 Partial Least Squares Regression

We use Partial Least Squares (PLS) Regression to build a model for perceiving a child's emotional state (represented by the two variables, Mood and Energy), given the vocal acoustic features described in Section 4.2. PLS regression is a multivariate statistical method for modeling a set of dependent variables in relation to a very large set of independent variables. Simply stated, PLS regression generalizes and combines concepts from ***Principal Components Analysis (PCA)*** and ***Multiple Regression*** (Abdi, 2003), performing an internal factor analysis designed to capture most of the information in the independent variables (input) that is useful for modeling the dependent variables (response) (Garthwaite, 1994).

The term *partial least squares* specifically means the computation of the optimal least squares fit to part of a correlation or covariance matrix (McIntosh et al., 2004; Pirouz, 2006; Wold, 1982). The part of the correlation or covariance matrix that the least squares are fit to is the "cross-block" correlation between the input variables and the response variables (Pirouz, 2006). PLS measures covariance between two or more blocks of variables and creates a new set of variables that is optimized for maximum covariance using the fewest dimensions (McIntosh et al., 1996; Pirouz, 2006). In this new set of variables, the input data matrix X and the response data matrix Y are modeled as linear combinations of a set of orthogonal factors, also known as latent variables, or components (De Vries & Ter Braak, 1995).

PLS is often compared to its closest alternative, Principal Components Regression (PCR). Like PCR, PLS regression takes as an input parameter the number of components to retain; when smaller than the feature space, this reduces dimensionality and guards against overfitting. In such comparative analyses, however, PLS Regression models consistently demonstrate superior classification accuracy and generalizability, with the smallest number of components (Garthwaite, 1994; The MathWorks; Yeniay & Goktas, 2002). This superior performance of PLS is attributed to the fact that PLS includes consideration of the response in addition to the inputs when forming these components, whereas PCR simply applies PCA internally to the inputs prior to regression. PLS is therefore considered especially useful for constructing prediction equations when there are many independent

variables and comparatively little sample data, particularly when the random error variance is large (Garthwaite, 1994; Hoskuldsson, 1988). However, even though there are no specific sample size requirements, the smaller the sample, the more likely that the model will be fitted to noise in data instead of the true distribution (Tobias, 1995).

A commonly used example in expositions of PLS regression involves a dataset from the field of chemometrics modeling that consists of 401 features and only 60 samples (Kalivas, 1997). The effectiveness of PLS regression when applied to this dataset, in terms of both completeness of the model and generalizability, speaks to its robustness against overfitting and small sample sizes (Garthwaite, 1994; The MathWorks; Tobias, 1995). With 82 features in our feature vector, and sample sizes that get quite small for certain subsets of the data (see Section 4.3.1), these characteristics make PLS regression an appealing method for our analysis.

4.3.1 Experimental Design

As part of our experimental design, we apply PLS regression to different subsets of the data, not only looking at the dataset as a whole, but also isolating more specific social and behavioral contexts based on agreed answers to the Social and Nature questions in the annotator questionnaire (Chapter 3). Table 4-3 lists these contexts and the number of samples corresponding to each, both for the overall time period, and per individual months' worth of data. The two dates in April 2007 are listed separately, because our intention for that period (15-24 months of age) is to study one day's worth of data per month. For the month of April, however, having first started with April 27, it became apparent that this day was unusual due to the new presence of the child's newborn sibling who had recently arrived from the hospital. For this reason, we collected an additional day's worth of data that month, April 9, and study it separately to control for any anomalous patterns that may be due to the extraordinary events on April 27.

Socio-Behavioral Context	Sample Size										
	Total	May 06	Jun 06	Jul 06	Aug 06	Sep 06	Oct 06	Jan 07	Apr 9 07	Apr 27 07	Jul 07
All Vocalizations	7376	527	495	911	965	741	843	625	1175	582	512
Social Situations	3857	284	187	522	501	337	562	367	638	251	208
All Nonbodily	4986	428	350	762	708	513	609	405	665	294	252
Social Nonbodily	3796	258	187	508	498	334	556	365	634	248	208
All Crying	398	70	22	109	48	59	25	23	27	3	12
Social Crying	336	41	17	103	47	40	25	21	27	3	12
All Laughing	57	4	3	10	8	8	2	2	7	11	2
Social Laughing	52	4	2	7	7	8	2	2	7	11	2
All Babble	656	4	19	26	103	33	67	53	144	120	87
Social Babble	506	3	8	17	63	23	59	47	130	96	60
All Speech	442	0	0	0	0	1	0	73	245	53	70
Social Speech	418	0	0	0	0	1	0	54	241	53	69
All Other Emoting	3433	350	306	617	549	412	515	254	242	107	81
Social Other Emoting	2484	210	160	381	381	262	470	241	229	85	65

Table 4-3. Sample sizes, per situational and monthly subsets of the dataset. Subsets highlighted in yellow were retained for analysis.

Subsets with fewer than 20 samples are left out of the analysis. For the Laughing contexts, this leaves out all of the monthly subsets, which eliminates Laughing from our longitudinal analysis. For each of the remaining 113 subsets, highlighted yellow in Table 4-3, we build five PLS Regression models, each distinguished by a different set of input variables:

(1) The 82 child vocalization features;

(2) The 82 features for adult speech ***before*** each vocalization;

(3) The 82 features for adult speech ***after*** each vocalization;

(4) The 82 features for adult speech ***surrounding*** each vocalization; and

(5) All of the above together, totaling 328 features.

Section 4.3.2 describes the procedures and parameters used to build each individual PLS Regression model.

The input response variables are the Mood and Energy ratings selected and agreed upon by annotators during the data collection phase (Chapter 3). We define agreement according to the 9-point scale described in Chapter 3, because this method produced the highest inter-annotator agreement by far in our evaluation of different scale variants. In the 9-point scale method, disagreements differing by one point are considered agreements with a value that is the average of the two ratings.

We evaluate each model for its accuracy in perceiving the child's emotional state using adjusted R-squared, as described in Section 4.3.2. We plot adjusted R-squared overall for the 9-24 month period as a function of the social and behavioral context. For each of these contexts, we also plot the same metrics longitudinally as a function of the ten regression models in the month-by-month timeline. Finally, we repeat both socio-behavioral and longitudinal analyses for specific child-caretaker dyads.

4.3.2 Parameters, Procedures, and Metrics

We compute PLS regression models for our analysis using Matlab's **`plsregress`** function, with 10-fold cross validation, which takes as input an input data matrix X, a response data matrix Y, and the number of PLS components to retain. The input dataset is normalized with the **`zscore`** function prior to input. As output, **`plsregress`** returns:

- **Factor loadings for X and Y** – the coefficients that specify the linear combinations comprising the internally constructed PLS components.

- **Scores** – the projections of the original input data onto the new space defined by the PLS components.

- **Regression coefficients B** that define the constructed model.

- **Mean-squared error (MSE)** – the variance in the test dataset that is unexplained by the model. By computing the difference between the responses predicted by the

model and the actual responses in the testing dataset for each observed set of input values, MSE evaluates a model's ability to predict new samples when applied to another dataset in the same feature space. An MSE closer to zero means better generalizability, and therefore, better accuracy when applied to a new set of input data points.

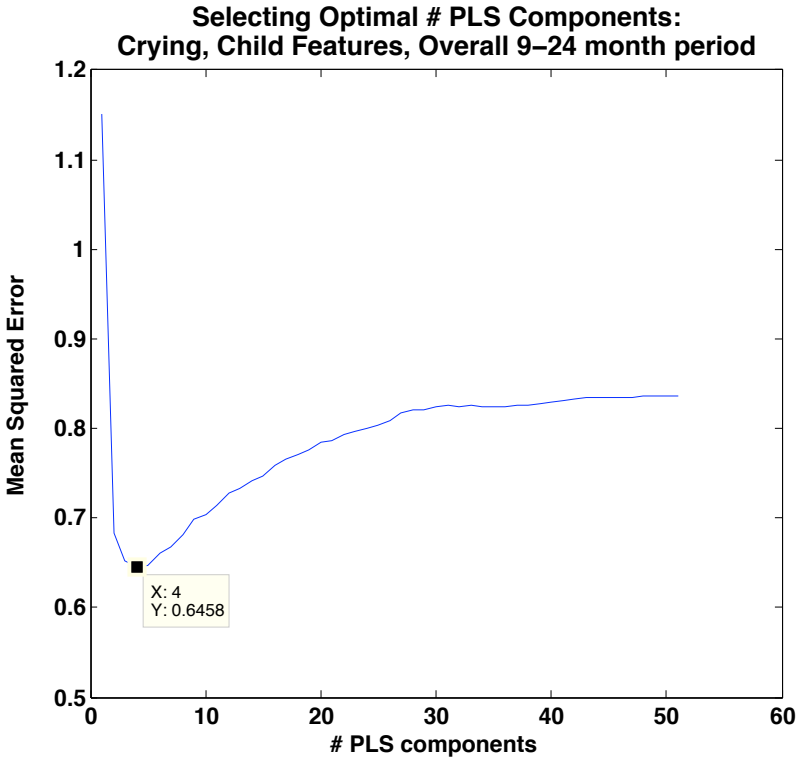


Figure 4-1. Deriving the optimal number of PLS components for a model. This particular example is for PLS regression applied to the Crying subset of child vocalization features, covering the overall 9-24 month period. A clear minimum MSE value occurs at 4 PLS components, beyond which MSE starts to rise. We therefore build the PLS model for this subset of data using 4 PLS components.

For each individual model, we derive the optimal number of PLS components to retain by first running **`plsregress`** using 50 components, plotting MSE as a function of number of components[12], and choosing the minimum (Hubert & Vanden Branden, 2003; Rosipal &

[12] Internally, **`plsregress`** does an iterative analysis of MSE, in which it computes MSE incrementally starting from just one component, and adding another component each time until all the components in the model are included. Thus, MSE is returned by **`plsregress`** in the form of a vector equal in length to the number of components, plus one, to account for a constant term in the model. Each element of this vector is the MSE that is computed for the number of components signified by its index in the vector.

Trejo, 2001; The MathWorks), as demonstrated by Figure 4-1. Appendix F lists the optimal number of components for each of our overall (9-24 month period) models.

For each individual model, we derive the optimal number of PLS components to retain by first running **`plsregress`** using 50 components, plotting MSE as a function of number of components[13], and choosing the minimum (Hubert & Vanden Branden, 2003; Rosipal & Trejo, 2001; The MathWorks), as demonstrated by Figure 4-1. Appendix F lists the optimal number of components for each of our overall (9-24 month period) models.

To evaluate the performance of each model and explore the correlations that it represents, we compute ***adjusted R-squared***, R^2_{adj}. First we externally compute ***R-squared***, a measure of the goodness of fit of the model[14], using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \left(\sum_{j=1}^{p} x_{ij}B_j\right)\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{SS_{err}}{SS_{tot}}$$

Equation 4-1. Formula for computing R-squared

where *n* is the number of samples; *p* is the number of input variables; Y_i is the vector of multivariate response values for sample *i;* B_j is the regression coefficient corresponding to feature *j* in the model; x_{ij} is the input value for feature *j* in sample *i;* and $\bar{Y}$ is the mean of the response values over all samples, a vector in which the mean for each dimension of the response is stored in a separate element.

The term $\left(\sum_{j=1}^{p} x_{ij}B_j\right)$ feeds the inputs x_{ij} into the model, computing the response that is predicted by the model, given the input predictor values. The numerator in the second term

[13] Internally, **`plsregress`** does an iterative analysis of MSE, in which it computes MSE incrementally starting from just one component, and adding another component each time until all the components in the model are included. Thus, MSE is returned by **`plsregress`** in the form of a vector equal in length to the number of components, plus one, to account for a constant term in the model. Each element of this vector is the MSE that is computed for the number of components signified by its index in the vector.

[14] http://en.wikipedia.org/wiki/Coefficient_of_determination

of Equation 4-1 is therefore computing the sum of squares of the differences between the ***actual*** response values Y_i and the ***predicted*** response values over all *n* samples. This value is SS_{err}, the residual sum of squares[15]. The denominator is computing the sum of squares of the differences between the actual response values and the sample mean, which is defined as SS_{tot}, the total sum of squares[16].

R-squared represents the proportion of variability in the training data that is accounted for by the regression model, or how well the regression line approximates the input data points. When R-squared is close to 1, this means a good fit; when close to 0, a poor fit. Because PLS regression is a form of linear regression, R-squared can also be equivalently defined and interpreted as the square of the sample correlation coefficient between the predictor data and the response data.

However, due to the statistical shrinkage effects known to occur in regression analysis[17], R-squared alone cannot be used as a meaningful comparison of models with varying numbers of explanatory terms and samples. We therefore adjust R-squared to obtain a metric that is normalized against both sample size and number of explanatory terms. In the context of PLS Regression, the number of explanatory terms in each model is the optimal number of PLS components that we select for that model, using the heuristic in Figure 4-1. We apply the following formula to compute adjusted R-squared (Harlow, 2005):

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-q-1} = 1 - \frac{MSE}{MST}$$

Equation 4-2. Formula for Computing Adjusted R-Squared

where R^2 is the R-squared value computed in Equation 4-1, ***n*** is the sample size, and ***q*** is the number of explanatory terms, or PLS components, used to build the model. The second term of Equation 4-2 is equivalently described as ratio of Mean Squared Error (MSE) (i.e, unexplained variance) and Mean Square for Treatments (MST) (i.e. explained variance), and is the inverse of the F-test statistic, where F = MST/MSE (Keller, 2009).

---

[15] http://en.wikipedia.org/wiki/Coefficient_of_determination

[16] http://en.wikipedia.org/wiki/Shrinkage_(statistics)

Effect size	Adjusted R-squared Range
Small	$0.02 \leq R^2_{adj} < 0.13$
Medium	$0.13 \leq R^2_{adj} < 0.26$
Large	$0.26 \leq R^2_{adj}$

Table 4-4. Effect Size Scale for Interpreting adjusted R-squared

We use adjusted R-squared to evaluate the effect size of our models (Cohen, 1992; Harlow, 2005). Because effect size is a measure of the strength of the relationship between input and response variables, this is directly related to R-squared as a measure of correlation, as well as the classification accuracy of a model. To interpret the effect size of a model given by its adjusted R-squared, we use the standard scale (Cohen, 1992; Harlow, 2005) listed in Table 4-4.

Chapter 5

PLS Regression Analysis Results

We have conducted an exploratory Partial Least Squares regression analysis to investigate the potential for creating a model that can simulate human perception of child and adult vocal expression to determine a child's emotional state. To implement our methodology, we applied the methods described in Chapter 4 to the data on the child's vocalizations and perceived emotional state that we collected in Chapter 3. In addition to modeling the dataset as a whole, we built separate PLS regression models that isolate specific socio-behavioral, dyadic, and longitudinal contexts. We examined the strength of the correlations that these models have captured between the child's perceived emotional state and both child and surrounding adult vocal acoustic features. Mapping these correlations enabled us to explore

- The degree to which caretaker speech reflects the emotional state of the child;
- How these correlations differ for specific child-caretaker dyads;
- Whether any stronger correlations emerge in certain social or behavioral situations (such as when the child is crying, babbling, or laughing); and
- Any longitudinal trends that might reveal a developmental progression of these correlations during the child's growth from 9 to 24 months of age.

Our analysis, and the results we present in this chapter, is structured in two parts. In the first part (Section 5.1), we evaluate PLS regression models across socio-behavioral contexts using the data from the entire 9 to 24 month period at once. The second part (Section 5.2) is a longitudinal study in which we build separate PLS models for each month of data and

track how their perceptual accuracy changes during the child's development from 9 to 24 months of age. We look at these longitudinal trends separately for each of the socio-behavioral contexts analyzed in Section 5.1.

Although orthogonal in the insights they reveal, both studies have a similar design as described in detail in Chapter 4 and summarized here for clarity. For each category (socio-behavioral contexts in the first study, months in the second study), we built separate PLS regression models for each of five feature sets:

(1) ***Child*** - the set of 82 features computed from the child's vocalizations

(2) ***Adult before*** – the set of 82 features computed from adult speech occurring within 30 seconds *before* each child vocalization

(3) ***Adult after*** – the set of 82 features computed from adult speech occurring within 30 seconds *after* each child vocalization

(4) ***Adult surrounding*** – the set of 82 features computed from the concatenation of (2) and (3); in other words, adult speech *surrounding* each child vocalization

(5) ***All of the above*** (or ***combined***) – the union of all four feature sets above, consisting of 82x4 = 328 features.

Each study consists of an overall and a dyadic component. The overall component represents all caretakers as a group. In computing the *adult before*, *adult after*, and *adult surrounding* feature sets, no distinction was made based on caretaker identity. In the dyadic component, we computed these feature sets separately for each caretaker (see Section 4.1.3), repeated the analysis for each child-caretaker dyad, and compared differences between them. Our dyadic analysis focuses on the three main caretakers of the child: the Father (Adult 0), the Nanny (Adult 1), and the Mother (Adult 2).

Our main metric in this analysis is adjusted R-squared, as described in Section 4.3. The many synonymous analytical interpretations of adjusted R-squared – such as the *effect size*,

the square of the *correlation coefficient*, *variance explained*, *goodness of fit*, and the likelihood of a model's *predictive accuracy* – make it a versatile metric that is well suited for our exploratory analysis. In parallel, we also plot the variance of the response variables (mood and energy) to account for any cases where adjusted R-squared might increase due to low variance in the response.

5.1 Adjusted R-squared Across Socio-Behavioral Contexts

In this section, we present the results of our first study, in which we built and evaluated PLS regression models for the overall 9 to 24 month time period, across the 14 socio-behavioral contexts listed in Table 4-3. Section 5.1.1 gives the results for all adults as a group, and Section 5.1.2 gives the dyad-specific results. As a brief overview, we make the following observations for discussion in Chapter 6:

- All adjusted R-squared values for both ***child*** and ***combined*** models qualify as very large effect sizes, with values much greater than the 0.26 cutoff given by Cohen (1992) and Harlow (2005). This strong model performance, suggesting high correlation, is quite consistent regardless of socio-behavioral context.

- Combining all ***child*** and adult-only feature sets yields only slightly better model performance than just using the child features.

- Adult-only models generally have much smaller adjusted R-squared than ***child*** or ***combined*** models. Even so, adult-only models achieve medium effect size in many contexts, including the overall dataset, all social situations and nonbodily vocalizations.

- The adult-only results reveal some interesting correlations that may point to common patterns of adult behavior in response to specific socio-behavioral contexts: correlations tend to be highest for crying, and lowest for babble and

speech. Social situations tend to bring out a slight increase in correlation between adult speech and perceived child emotion. Adult speech ***after*** a child's Social Laughing vocalizations is significantly more correlated than ***before*** or ***surrounding***.

- Adult-aggregate models (Section 5.1.1), which represent all caretakers as a group, tend to have equal or higher adjusted R-squared values than dyad-specific models (Section 5.1.2).

- Some notable caretaker-specific patterns emerge in the dyadic analysis. In particular, the Nanny generally tends to have the highest adjusted R-squared among the caretakers. Among the more specific contexts, the Nanny is most correlated in Crying contexts, the Mother has the highest adjusted R-squared in Laughing contexts, and the Father is highest correlated in Social Babble and Social Speech.

5.1.1 All Adults

Figure 5-1 shows adjusted R-squared as a function of socio-behavioral context, for child, adult, and combined[18] PLS models computed for all time and all caretakers in aggregate. The variances for each of the two response variables, mood and energy, are also included for reference. The best performing PLS models are those built using the ***combined*** feature set that includes all four child and surrounding adult feature sets. Here, adjusted R-squared ranges from 0.54 to 0.6, with an average of 0.58, with marked consistency across socio-behavioral contexts. PLS models for the child also indicate high correlations (range: 0.5 to 0.57, average: 0.536), although they are slightly smaller than for the combined feature set. Like the ***combined*** models, the ***child*** models are consistent across contexts, with the exception of Laughing and Social Laughing. Here, adjusted R-squared for the child drops to 0.34 and 0.35, respectively. We note that all adjusted R-squared values for both ***child*** and ***combined*** models qualify as very large effect sizes, with values much greater than the 0.26 cutoff given by Cohen (1992) and Harlow (2005).

[18] Henceforth, we use ***all of the above*** and ***combined*** interchangeably to describe the feature set that takes the union of the child, adult before, adult after, and adult surrounding feature sets.

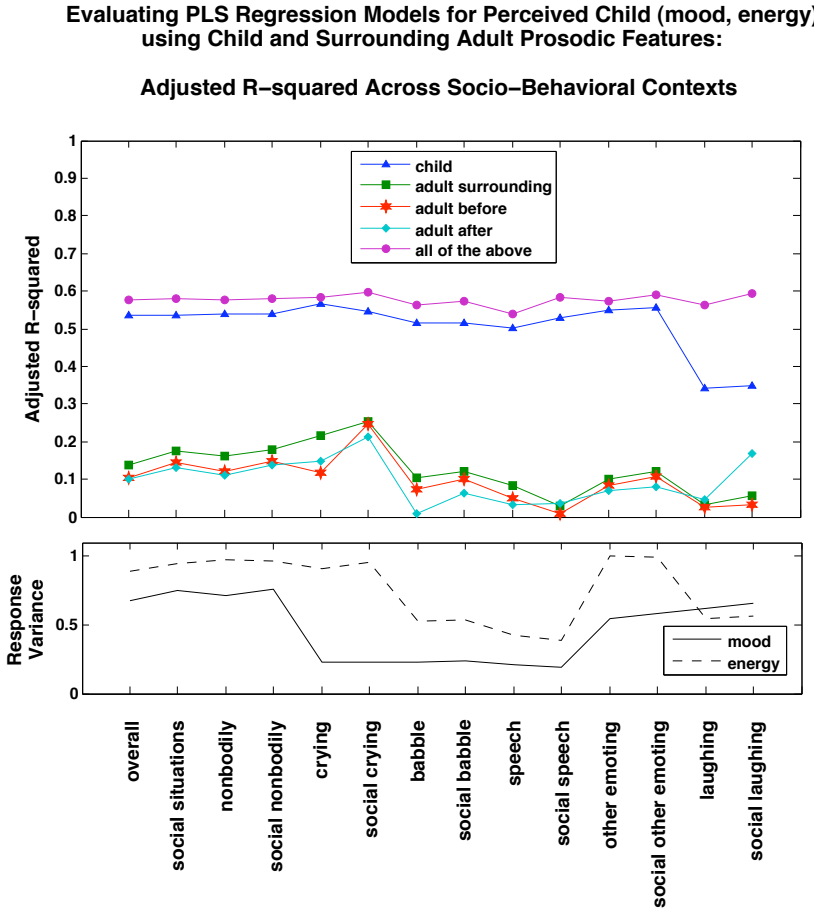


Figure 5-1. Adjusted R-squared and Response Variance across Socio-Behavioral Contexts, for all time and all caretakers in aggregate.

Compared to the ***child*** and ***combined*** models, PLS models built using adult-only feature sets maintain a significantly smaller adjusted R-squared, ranging between negligible (less than 0.02) to a high of 0.25. Average adjusted R-squared is 0.13 for ***adult surrounding***, 0.097 for ***adult before***, and 0.096 for ***adult after***. Particularly interesting, however, is that for many contextual subsets, including the overall dataset as a whole, the adult-only models achieve an adjusted R-squared greater than 0.13, which is interpreted as a medium effect size by Cohen and Harlow. Table 5-1 lists these cases together with their adjusted R-squared values for each of the adult-only models. Of special note among these results are the Crying contexts, both overall and in social situations, which elicit the highest adjusted R-squared performance among adult-only models. However, we note the drop in

mood variance during crying, which may partially explain the rise in adjusted R-squared. We discuss this point further in Chapter 6.

Adult Surrounding		Adult Before		Adult After	
Context	R^2_{adj}	Context	R^2_{adj}	Context	R^2_{adj}
Overall	0.138	All Social Situations	0.144	All Social Situations	0.133
Social Situations	0.176	Social Nonbodily	0.147	Social Nonbodily	0.136
All Nonbodily	0.162	Social Crying	0.247	All Crying	0.146
Social Nonbodily	0.178			Social Crying	0.212
All Crying	0.216			Social Laughing	0.168
Social Crying	0.255				

Table 5-1. Adjusted R-squared of Socio-Behavioral Contexts Eliciting Medium Effect Size in Adult-Only PLS Regression Models. This is notable, because it suggests that adult speech surrounding a child's vocalization can tell us something meaningful about the child's emotional state.

Also interesting among the adult-only results is the performance of the ***adult after*** model in the Social Laughing context. It achieves a significantly higher adjusted R-squared than either ***adult before*** or ***adult surrounding***, even qualifying as a medium effect size, with a value of 0.168. This result may reflect a pattern of caretakers laughing in response to the child's laughter, which would seem to be a natural, common occurrence in a social context.

Adjusted R-squared drops significantly for Babble and Speech in the adult-only models, even becoming negligible for Social Speech. With this one exception, social situations across behavioral contexts tend to bring out a slight increase in correlation between adult speech and child emotion.

5.1.2 Dyadic Analysis

In Figure 5-2, we compare adjusted R-squared trends between dyads, for each adult-based feature set separately, across socio-behavioral contexts. Each graph in Figure 5-2 corresponds to one of the four adult-related feature sets – ***adult before***, ***adult after***, ***adult surrounding***, and ***combined*** – and plots adjusted R-squared as a function of context for the Father, Nanny, and Mother. We also include the corresponding overall trend from Section

5.1.1, Figure 5-1 as a dotted line, to compare how PLS models for each individual adult compare with the aggregate PLS models representing all adults as a group. As a general pattern, we observe that, with just a few exceptions, the aggregate models representing all caretakers as a group tend to have equal or higher adjusted R-squared values than the dyad-specific models.

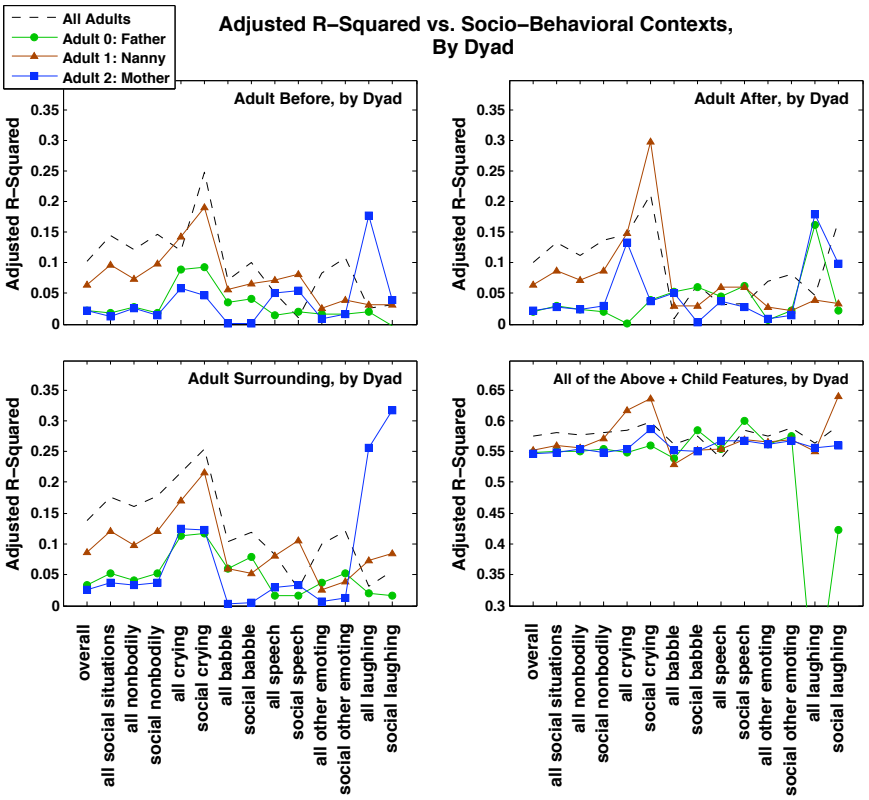


Figure 5-2. Dyadic Analysis of Adult-Specific PLS Models across Socio-Behavioral Contexts.

Among the three dyads, however, adjusted R-squared tends to be highest overall for the Nanny, indicating a consistently higher correlation between the Nanny's speech and the child's emotions than for the other caretakers. Within this overall trend, we note in particular the Social Crying context, in which the Nanny's models consistently score the highest across feature sets among the caretakers: 0.191 using ***adult before***, 0.297 using ***adult after***, 0.217 using ***adult surrounding***, and 0.636 using the ***combined*** feature sets. Adjusted R-squared for All Crying is also consistently higher in all the Nanny's models

relative to the other dyads, with values of 0.142 for ***adult before***, 0.148 for ***adult after***, 0.170 for ***adult surrounding***, and 0.616 for ***combined***.

In certain other contexts, such as Laughing, the Mother has significantly higher correlation than other caretakers, as seen by the adjusted R-squared values of 0.256 and 0.318 for All Laughing and Social Laughing, respectively, that are achieved by her ***adult surrounding*** models. The latter, 0.318 for Social Laughing, falls into the large effect size scale, and is, in fact, the highest value achieved by any adult-only model in this part of the analysis. In the ***combined*** feature set, however, it is the Nanny who achieves the highest adjusted R-squared in Social Laughing, with a value of 0.641, compared to 0.561 for the Mother and 0.423 for the Father.

Adjusted R-squared for the Father tends to slightly exceed the other caretakers in Social Babble and Social Speech contexts, with values of 0.584 and 0.599 in the corresponding ***combined*** models, which even surpass the adult-aggregate models for these contexts (values: 0.5744 and 0.585, respectively).

Beyond these differences, we observe that the Mother and Father have fairly similar patterns of correlation across socio-behavioral contexts. Where they diverge, it is in most cases because of greater similarity to the Nanny, such as for All Babble in ***adult surrounding*** for the Father, All Crying in ***adult after*** for the Mother, and All Laughing in ***combined*** for the Mother. In any case, given the zoomed-in scale shown in Figure 5-2, all differences beyond those already mentioned are negligible.

5.2 Longitudinal Analysis

In this section, we present the results of the longitudinal analysis, in which we build and evaluate month-specific PLS regression models, and investigate any progressive trends over time that may indicate a developmental trajectory. In Section 5.2.1, we describe the results for all adults as a group. Section 5.1.2 briefly addresses the dyad-specific results. Our observations from this longitudinal analysis are summarized as follows:

- For the most general contexts – All Vocalizations, Social Situations Only, All Nonbodily Vocalizations, and Social Nonbodily Vocalizations – the trends are flat, showing no consistent, progressive increases or decreases in correlation (and thus, perceptual accuracy) over time. These four contexts produce very similar longitudinal graphs.

- Month-specific models tend to have significantly higher adjusted R-squared than time-aggregate models, both individually, and on average.

- Month-specific ***combined*** models tend to improve upon their corresponding ***child*** models to a greater degree than do time-aggregate models.

- We observe notable longitudinal progressions in adjusted R-squared within the Crying, Babble, and Other Emoting contexts: increasing over time for Crying and Other Emoting, and decreasing for Babble. Child and combined models achieve very high values of adjusted R-squared during these progressions, above 0.80.

- No notable dyad-specific patterns emerge beyond the confirmation that adult-aggregate models tend to be consistently more accurate than, or at least equal in performance to, each adult individually.

5.2.1 All adults

Figure 5-3 shows the longitudinal adjusted R-squared trends for the four most general socio-behavioral contexts in our analysis: All Vocalizations, Social Situations Only, All Nonbodily Vocalizations, and Social Nonbodily Vocalizations. We present them here as a group, because all four graphs bear a clear similarity to each other. Our first observation is that these trends are flat, showing no consistent, progressive increases or decreases in correlation (and thus, perceptual accuracy) over time. Adjusted R-squared is roughly within the same scale of variance at 15-24 months as at 9 months and remains within this

range throughout the period of analysis. However, in several more specific contexts, such as Crying, Babble, and Other Emoting, we do observe notable longitudinal changes in correlation, which we describe later in this section.

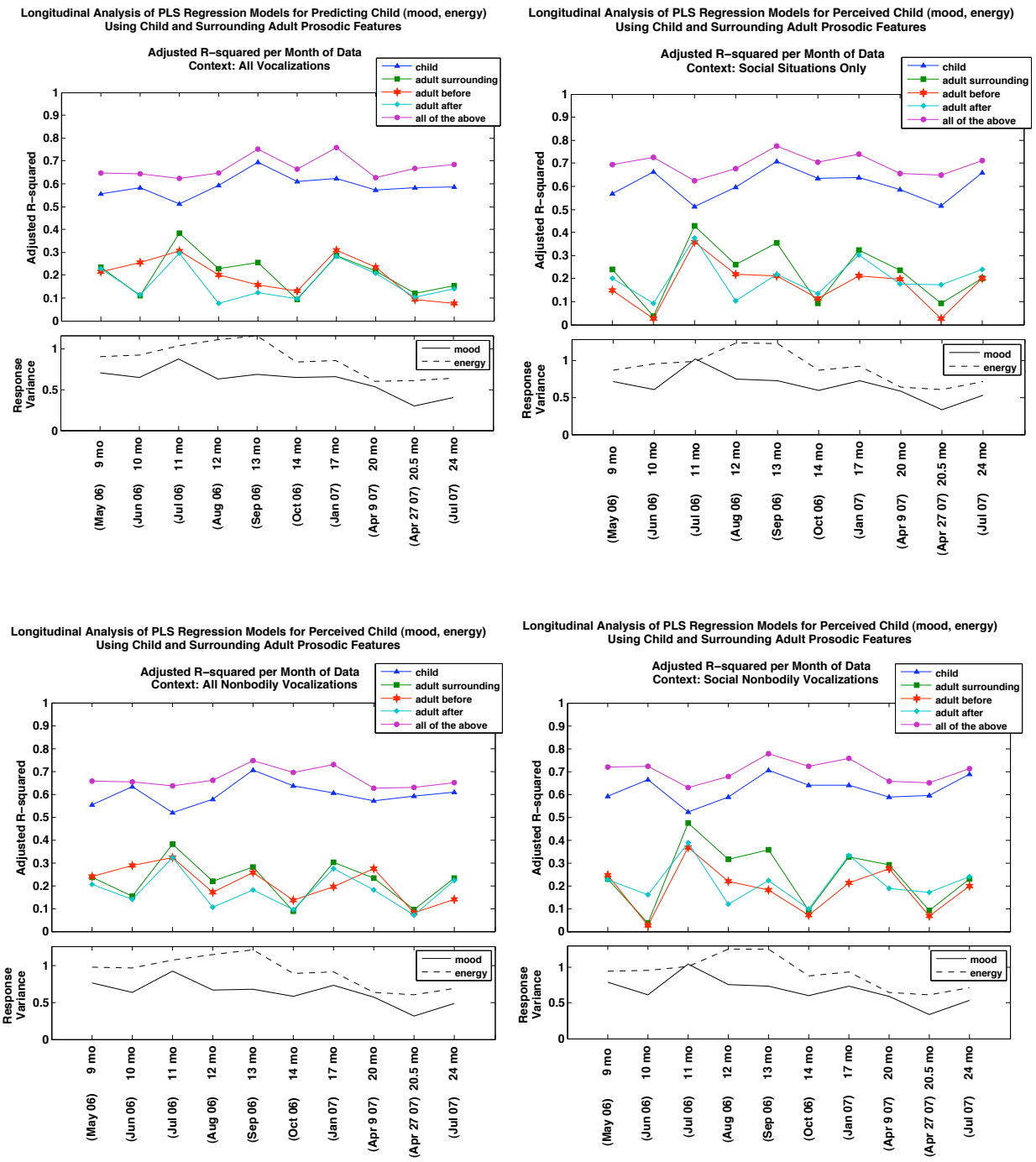


Figure 5-3. Longitudinal Trends for All, Social Only, Nonbodily, and Social Nonbodily Vocalizations

Our second observation from the graphs in Figure 5-3 is that adjusted R-squared for the month-specific models is consistently higher than their time-aggregate counterparts that were analyzed in Section 5.1 (Figure 5-1). To illustrate this pattern more clearly, we present in Table 5-2 a comparison between the adjusted R-squared values of each time-aggregate model and the mean, minimum, and maximum adjusted R-squared of its corresponding month-specific models. Table 5-2 shows these statistics for the ***combined***, ***child***, and ***adult surrounding*** feature sets.

Table 5-2 demonstrates that this pattern holds consistently for all socio-behavioral contexts and all feature sets: the mean adjusted R-squared of month-specific models is higher than the adjusted R-squared of the corresponding time-aggregate models. For the most general context, All Vocalizations, the time-aggregate ***combined*** model has an adjusted R-squared of 0.58, while the month-specific ***combined*** models average out to 0.67 (range: 0.62 to 0.76), a significant increase in correlation and perceptual accuracy. Similarly, the time-aggregate ***child*** model has an adjusted R-squared of 0.54, while the month-specific mean is 0.59, with a range of 0.51 to 0.69. This holds even for adult-only models: the time-aggregate R-squared for ***adult surrounding*** is 0.14 in the All Vocalizations context, compared to a monthly mean of 0.21 (range: 0.09 to 0.38). Although not included in Table 5-2, the ***adult before*** and ***adult after*** models are consistent in this pattern as well.

Context	**Combined**				**Child**				**Adult Surrounding**			
	Overall	Longitudinal			Overall	Longitudinal			Overall	Longitudinal		
	(Fig 5-1)	mean	min	max	(Fig 5-1)	mean	Min	max	(Fig 5-1)	mean	min	max
All Vocalizations	0.58	0.67	0.62	0.76	0.54	0.59	0.51	0.69	0.14	0.21	0.09	0.38
Social Only	0.58	0.70	0.62	0.78	0.54	0.61	0.51	0.71	0.18	0.23	0.04	0.43
All Nonbodily	0.58	0.67	0.63	0.75	0.54	0.60	0.52	0.70	0.16	0.22	0.09	0.38
Social Nonbodily	0.58	0.70	0.63	0.78	0.54	0.62	0.52	0.71	0.18	0.24	0.04	0.48
All Crying	0.59	0.78	0.31	1.00	0.57	0.76	0.61	0.97	0.22	0.40	0.09	1.00
Social Crying	0.60	0.86	0.71	1.00	0.55	0.74	0.61	0.98	0.26	0.51	0.13	0.81
All Babble	0.56	0.73	0.56	0.98	0.52	0.68	0.54	0.88	0.10	0.32	0.01	0.79
Social Babble	0.57	0.82	0.60	0.99	0.52	0.66	0.53	0.84	0.12	0.26	0.00	0.60
All Speech	0.54	0.71	0.58	0.87	0.50	0.63	0.49	0.78	0.08	0.18	0.06	0.39
Social Speech	0.59	0.73	0.57	0.95	0.53	0.64	0.48	0.80	0.03	0.18	0.04	0.39
All Other Emoting	0.57	0.70	0.54	0.78	0.55	0.64	0.47	0.77	0.10	0.21	0.09	0.35
Soc Other Emoting	0.59	0.75	0.55	0.92	0.56	0.66	0.43	0.78	0.12	0.23	0.03	0.38

Table 5-2. Comparing Overall Performance of Month-by-Month models with Time-Aggregate models. For each Socio-Behavioral context, adjusted R-squared is significantly higher on average when using month-specific models than when using a time-aggregate model built using data from the overall 9-24 month period.

Our third observation from Figure 5-3 is that month-specific ***combined*** models for these four contexts tend to improve upon their corresponding ***child*** models to a greater degree than do time-aggregate models. In the All Vocalizations context, we note improvement by as much as 0.014 in the Jan 2007 data. On average, the month-specific ***combined*** models among the four contexts in Figure 5-3 outperform the ***child*** models by 0.08 adjusted R-squared, with a minimum improvement of 0.06. This is in contrast to the difference of 0.04 in the time-aggregate case (Figure 5-1).

Despite the longitudinal consistency of adjusted R-squared in Figure 5-3, we observe notable longitudinal progressions in the Crying, Babble, and Other Emoting contexts. As mentioned in Chapter 4, we exclude Laughing from longitudinal analysis due to monthly sample sizes less than 20 across the board. Also, due to sparseness of Speech vocalizations in the first six months of our longitudinal timeline (see Table 4-3), there are not enough data points in the graphs for All Speech and Social Speech to reveal any meaningful longitudinal trends. The longitudinal graphs for All Speech and Social Speech are included in Appendix G.

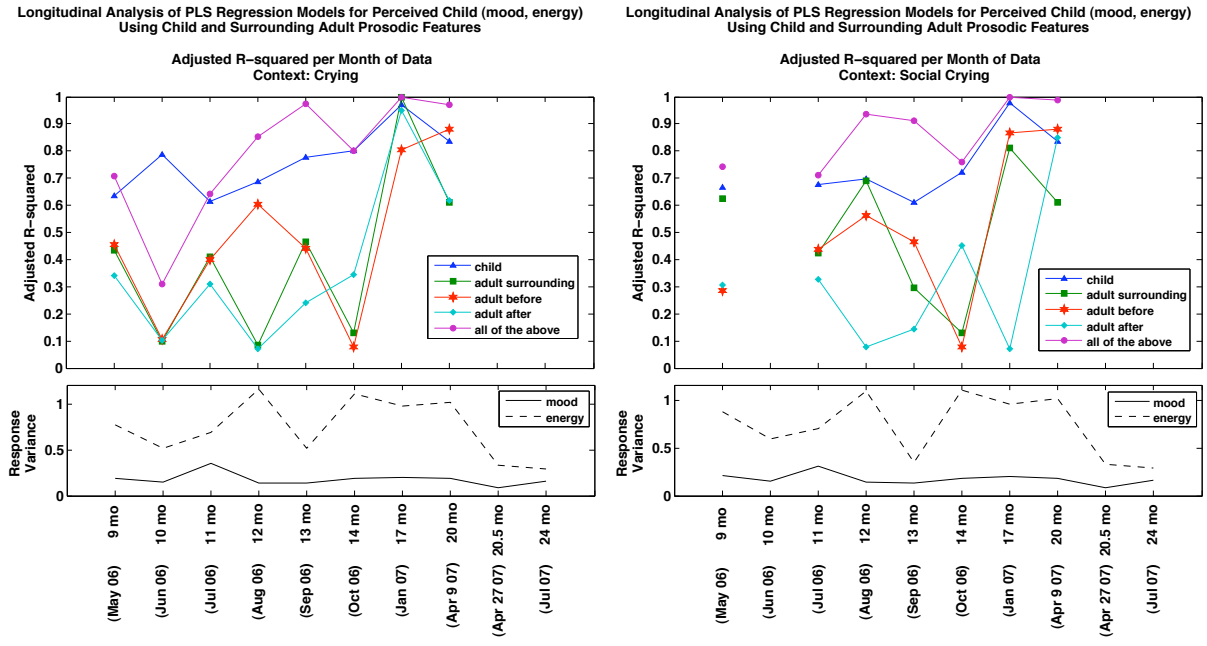


Figure 5-4. Longitudinal Trends in Adjusted R-squared for All Crying (left) and Social Crying (right).

The longitudinal trends for Crying and Social Crying in Figure 5-4 show an increasing progression in correlation (and perceptual accuracy) over time for the ***child*** and ***combined*** models. Also, correlations indicated by adjusted R-squared are generally much larger than in the more general contexts of Figure 5-3, with an average of 0.78 (Crying) and 0.86 (Social Crying) for ***combined*** and 0.76 (Crying) and 0.74 (Social Crying) for ***child***.

Socio-Behavioral Context	Sample Size										
	Total	May 06	Jun 06	Jul 06	Aug 06	Sep 06	Oct 06	Jan 07	Apr 9 07	Apr 27 07	Jul 07
All Crying	398	70	22	109	48	59	25	23	27	3	12
Social Crying	336	41	17	103	47	40	25	21	27	3	12

Table 5-3. Total and Monthly Sample Sizes for the All Crying and Social Crying Contexts.

Although the steadily increasing longitudinal progression for the child seems to be a clear, consistent trend for both All Crying and Social Crying, we note the particularly small sample sizes (between 21 and 27) for Jun 2006, Oct 2006, Jan 2007, and Apr 9 2007, listed in Table 5-3. Based on results in our other graphs in this work that include small sample sizes (see Figure 5-5, for example), there appear to be no consistent patterns indicating that sample size is an issue. We do concede, however, that using 328 features in the combined feature set (and potentially even 82 features) for just 20 or so samples may be reaching the limits of PLS regression's robustness against overfitting. For this reason, we reserve any final conclusions about the progressions in Figure 5-4, and the strength of the correlations that occur, pending further analysis with more samples.

In contrast to the increasing trends for Crying, Figure 5-5 shows a clear theme of downward longitudinal trends within the Babble contexts. The three most striking, compelling progressions occur in All Babble, for the ***child***, ***adult surrounding***, and ***combined*** feature sets. For ***child***, adjusted R-squared begins at 0.88 in Jun 2006, at 10 months of age, and decreases in a zig-zag manner to a low of 0.54 during Apr 9 2007, at 20 months of age, remaining at 0.60 and 0.62 thereafter. For ***adult surrounding***, adjusted R-

squared decreases in a nearly linear progression starting at 0.79 in Jun 2006 and ending at near zero, with 0.06 on Apr 9 2007 and 0.02 in Jul 2007[19].

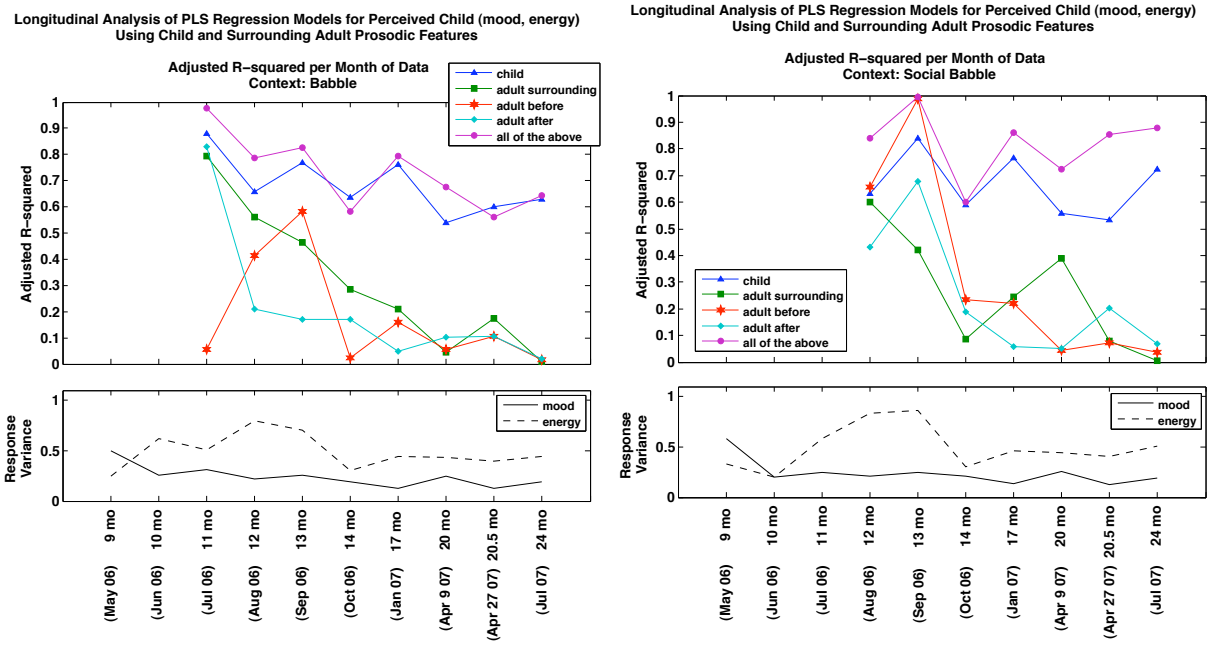


Figure 5-5. Longitudinal Trends in Adjusted R-squared for All Babble (left) and Social Babble (right).

Socio-Behavioral Context	Sample Size										
	Total	May 06	Jun 06	Jul 06	Aug 06	Sep 06	Oct 06	Jan 07	Apr 9 07	Apr 27 07	Jul 07
All Babble	656	4	19	26	103	33	67	53	144	120	87
Social Babble	506	3	8	17	63	23	59	47	130	96	60

Table 5-4. Total and Monthly Sample Sizes for All Babble and Social Babble contexts.

Adjusted R-squared for ***combined*** mirrors the zig-zag shape of the ***child*** trend, but has a slightly steeper downward slope, starting at 0.98 in June 2006 (10 months of age) and ending at 0.64 in Jul 2007 (24 months of age). For Social Babble, the most striking trends are for ***adult before*** and ***adult after***. The ***child*** and ***combined*** trends are flat, however, showing none of the decreasing slope seen in the All Babble context.

[19] The apparent anomaly at Apr 27 2007, may be explained by the unusual circumstances of that day, as noted in Chapters 3 and 4.

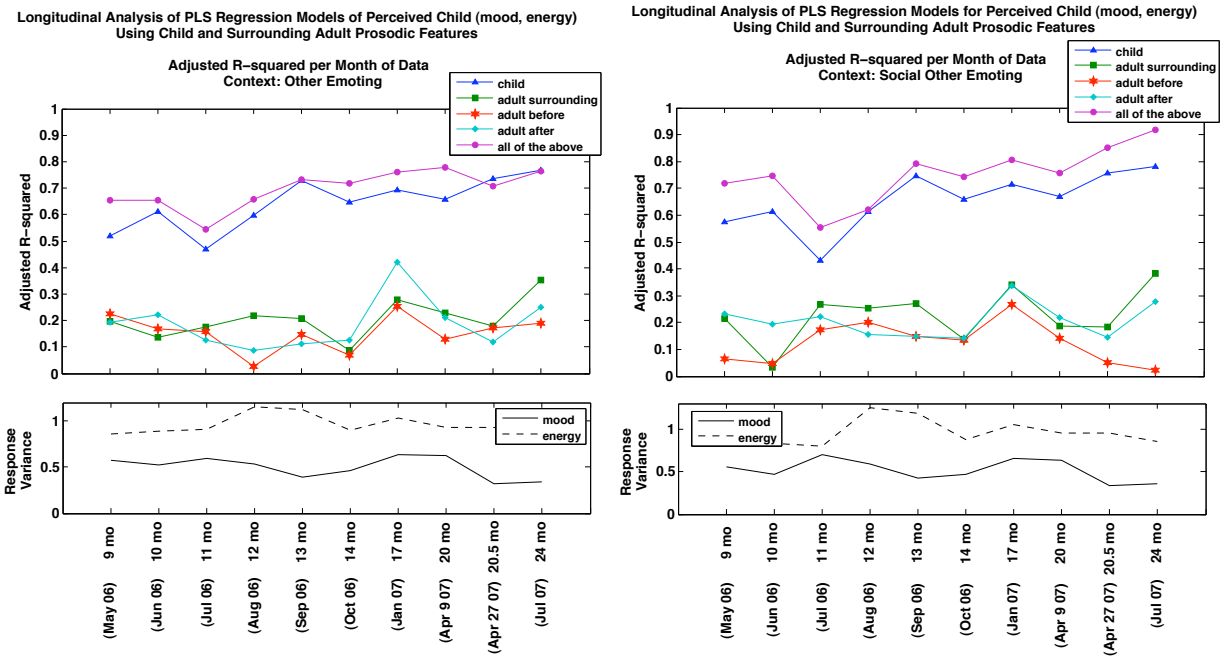


Figure 5-6. Longitudinal Trends in Adjusted R-squared for All Other Emoting (left) and Social Other Emoting (right)

It is in the context of Other Emoting that we observe what is perhaps the most interesting longitudinal progression in our results, where adjusted R-squared for ***child*** and ***combined*** increase steadily over time. As demonstrated by Figure 5-6, we see this progression in both All Other Emoting and Social Other Emoting contexts, with the latter demonstrating a slightly steeper upward slope for ***combined***. In All Other Emoting, the ***child*** trend starts with adjusted R-squared values of 0.52 at 9 months of age, in May 2006, and 0.47 at 11 months, in Jul 2006, but gradually rises over time, through values within the 0.60 to 0.69 range, ending with 0.74 at 20.5 months of age (Apr 27 2007) and 0.77 at 24 months (Jul 2006). The ***combined*** feature set starts higher, with 0.65 at 9 months of age (May 2006), but drops to a low of 0.54 at 11 months (Jul 2006) and increases progressively to end with 0.76 in Jul 2007, at 24 months of age. Along the way, the ***combined*** models consistently produce adjusted R-squared values above 0.70 starting at 13 months (Sep 2006), reaching a high of 0.78 at 20 months (Apr 9 2007).

The trends for Social Other Emoting are similar, with slight amplifications. Most notably, the ***combined*** feature set shows appreciably higher adjusted R-squared values across the board than for All Other Emoting, starting with 0.71 and 0.75 in May and Jun

2006 (9 and 10 months of age), respectively, then falling to a low of 0.55 at 11 months in Jul 2006, and increasing steadily thereafter, reaching into the 0.80 and 0.85 range and ending at a high of 0.92 at 24 months of age, in Jul 2007. The significance of these clear upward progressions, developing over time into very high correlations above 0.80, is supported by the fact that the monthly sample sizes for Other Emoting (shown in Table 5-5) are all decisively large enough to be practicable in PLS regression analysis.

Socio-Behavioral Context	Sample Size										
	Total	May 06	Jun 06	Jul 06	Aug 06	Sep 06	Oct 06	Jan 07	Apr 9 07	Apr 27 07	Jul 07
All Other Emoting	3433	350	306	617	549	412	515	254	242	107	81
Social Other Emoting	2484	210	160	381	381	262	470	241	229	85	65

Table 5-5. Total and Monthly Sample Sizes for All Other Emoting and Social Other Emoting Contexts.

5.2.2 Dyadic Analysis

In our analysis, we have also explored longitudinal trends between dyads. No notable dyad-specific patterns emerged, beyond the confirmation that adult-aggregate models tend to be consistently more accurate than, or at least equal in performance to, each adult individually, as we observed in Section 5.1.2. There are some occasional exceptions to this rule, such as the Father in the Speech contexts on Apr 27 2007 and in Jul 2007. However, we note that these exceptions, and the longitudinal variations between dyads more generally, may be due to some caretakers being more present than others during that month's worth of annotated data. Interpretation of these results may benefit from accounting for the degree to which each adult was present during each month's worth of data; we do not compute this here, but leave it as an indication for future extensions of this analysis. The graphs for the dyadic longitudinal analysis are included for reference in Appendix H.

Chapter 6

Discussion and Conclusions

In this thesis, we have developed and applied a methodology for modeling vocal acoustic correlates of the child's perceived emotional state within the Speechome corpus. To evaluate the potential of this methodology for creating an effective emotion recognition mechanism to support empirical study of early emotional processes in the Speechome corpus, we have conducted an exploratory analysis in Chapter 5 that compares child-only, adult-only, and combined models across socio-behavioral, dyadic, and longitudinal subsets of the data. Our experimental designs in this investigation have enabled us to evaluate the degree of perceptual accuracy that our modeling methodology achieves for a young child aged 9 to 24 months, as well as any methodological parameters that may be useful for optimizing perceptual accuracy. In addition, mapping the accuracy of these models across contexts has brought forth some preliminary insights about the vocal development of a child's emotional expression during this prelinguistic age period, en route to language. Quantifying the accuracy of adult-only models has also given us a rough measure of intersubjectivity, revealing developmental trends and caretaker-specific differences in the course of longitudinal and dyadic analysis.

6.1 Building a Perceptual Model for Child Emotion

Our results show great potential for automating the detection of a child's perceived emotional state in the Speechome corpus, by using child and adult vocal acoustic features. In Chapter 5, we observed consistently high adjusted R-squared values for Partial Least Squares (PLS) regression models built using only ***child*** features, as well as using a ***combined*** feature set that includes the adult feature sets along with these child features.

We are particularly encouraged by the consistency of these results, not only across specific socio-behavioral contexts, but also in the overall dataset, with the All Vocalizations context yielding adjusted R-squared values such as 0.54 and 0.59 for time-aggregate ***child*** and ***combined*** models, respectively, and 0.59 and 0.67 on average for corresponding monthly models. Such high adjusted R-squared values put the significance of these models, and the strength of the correlations they represent, well within the large effect size range, which is defined as 0.26 or greater.

Consistency across socio-behavioral contexts suggests that it may be unnecessary to build a specific classifier for distinguishing between them to capture optimal perceptual accuracy, simplifying the task of automating the perception of child emotions from naturalistic audio. In particular, this suggests that we may be able to eliminate the Social and Nature questions from our annotation methodology in future applications. This is an encouraging simplification, because the inter-annotator agreement for these questions was rather low, with Cohen's kappas of 0.552 and 0.499, respectively. On the other hand, the low agreement might itself suggest an ambiguity in our design of the Social and Nature questions that caused answers to these two questions to be unreliable. Given such low agreement, there is a possibility that annotators missed distinctions between certain contexts that may have been salient in our investigation.

This consistency also raises the possibility that we could build an overall model that would be consistently accurate in its perceptions regardless of the nature of the vocalization or whether it occurs during a social situation. We observe such consistency in several forms among our results, not only in the stability of adjusted R-squared values across contexts, but also in the similarity between longitudinal trends for All Vocalizations, Social Situations Only, Nonbodily Vocalizations, and Social Nonbodily Vocalizations. Although Social subsets of each behavioral context tend to bring out a very slight increase in adjusted R-squared, these qualities suggest that there is not much to gain by detecting social situations or filtering out bodily vocalizations, towards building a perceptual model of a child's emotions.

At first glance, Crying seems to be one exception to this rule, producing the highest adjusted R-squared values among the socio-behavioral contexts. In the monthly models, improvements in perceptual accuracy over other contexts range between 0.1 and 0.2 *on*

average, even achieving near-perfect correlation in some months. This high correlation is not surprising, since Crying is a very specific vocal behavior with characteristic acoustic properties that have been used to build robust cry detectors (Ruvolo & Movellan, 2008). Crying is also a context in which adults tend to respond with a relatively predictable repertoire of speaking styles characteristic of comforting the child. Interestingly, we note that the All Crying context is highest correlated in the ***child*** models, while Social Crying is highest correlated for ***combined*** and adult-only models. This pattern is just as one would expect, with models that involve adult features correlating more strongly within social situations, because it is in a social situation that an adult reacts directly to comfort a crying child.

However, we discount the significance of this superior recognition accuracy within Crying contexts, as well as its pertinence towards building a technology for perceiving child emotion in naturalistic audio recordings. Crying is a special case in that mood variance is restricted to less than half of the total rating scale; by definition, a young child cries only when his mood carries a negative valence. As mentioned above, there is also relatively low variability in adult response behavior to a child crying. Less variance to capture makes the classification task simpler and much more likely to succeed. What is most important to consider, however, is whether it is useful or meaningful to be able to detect a child's perceived emotional state if we already know that the child is crying. Beyond detecting subtle nuances within the state of negative valence (Gustafson & Green, 1989), this would seem to be a trivial task with little benefit.

On a similar note, we ask ourselves the question: is it worth computing feature sets for surrounding adult speech and including them in creating a perceptual model of child emotion? Judging by time-aggregate models alone, in which ***combined*** models introduce only a slight improvement over ***child*** models[20], the answer would be no. However, on a monthly basis, we observe that ***combined*** models introduce significant improvement over ***child*** models, which suggests that, on a per-month scale, adult data may be worth including in order to maximize the overall perceptual accuracy of our model.

[20] We treat the low perceptual accuracy of the time-aggregate child models in Laughing as an uncharacteristic anomaly that may be due to the small sample size for this context, and we therefore leave it out of the discussion.

This point is especially relevant, considering our key observation that monthly models tend to significantly outperform time-aggregate models on average: mean adjusted R-squared for monthly models is significantly higher than the adjusted R-squared value of the time-aggregate model for the corresponding socio-behavioral context. Besides its developmental implications, which we discuss in Section 6.2, this finding suggests that it is better to build separate models for each month of data than to have a single model for the entire dataset as a whole.

In light of these results, our recommendation for building a perceptual model for child emotion in applying Speechome for the study of emotional processes in early development is to use a feature set that combines child and adult acoustic vocal features and have separate models per month of data. In addition to this configuration, we add one more parametric recommendation: the dyadic analysis shows us that adult-aggregate models tend to be perceptually better than, or at least equal to, each adult individually. It is therefore preferable to build a single model using all adults than to build specific models for each adult, obviating the need for person-specific speaker identification to obtain optimal perceptual accuracy. However, a mechanism for distinguishing child from adult speech would still be in order.

Our results hold much promise towards building a fully automated technology for perceiving child emotion in naturalistic audio, but there is still some work to be done to achieve this vision. The high adjusted R-squared values that we observe in this work still depend on knowing the exact boundaries of a child's vocalizations, unobscured by noise or overlapping adult speech. As a next step, we recommend repeating this study using raw speech segments identified as child-produced by Speaker ID instead of our meticulously annotated child vocalization intervals. In this thesis, we have established a baseline by which we can evaluate the performance of this less stringent and more highly automated definition of child vocalizations. If overlapping adult speech and other noise proves to be detrimental to perceptual accuracy, a robust filter for removing overlapping adult speech and other noise from a child vocalization using spectral analysis and other signal processing methods could implement the missing link.

6.2 Developmental Insights

In addition to building a high-performing perceptual model, our results also demonstrate the potential of our methodology to reveal developmental insights into perceived child emotion and adult-child intersubjectivity. Although the work of this thesis is effectively a case study involving data for a single child, our longitudinal analysis has brought forth several interesting observations that seem to hint at developmental phenomena, inviting further validation by applying this methodology across multiple children. We recap these observations here and offer some hypotheses to explain their developmental significance.

The flat longitudinal trends for the most general contexts (All Vocalizations, Social Situations Only, All Nonbodily Vocalizations, and Social Nonbodily Vocalizations) seem to suggest that, in the general daily life experience of the child, the correlation between the child's perceived emotional states and the acoustic features of the child's vocalizations stays consistent over time, and is not an attribute that is subject to developmental change at this age. The same can also be said of surrounding adult speech – in that the correlation with child emotion stays fairly constant in the general case – indicating a consistency in the quality of intersubjectivity between child and caretakers during this period of development.

While the correlation itself may remain consistent over time, our results also demonstrate that the particular set of acoustic vocal features that best correlate with child emotion evolves with development, without changing the correlation itself. Our finding that monthly models are more successful than time-aggregate models seems to suggest this kind of developmental shift, such that looking at the time period as a whole will dilute those correlations. We explain this as an artifact of accelerated physiological changes in vocal tract anatomy (Kent & Murray, 1982; Vorperian et al., 2005), as well as cognitive development in the neural mechanisms for motor control of speech production (Petitto & Marentette, 1991; Warlaumont et al., 2010) that occur during infancy and early childhood. Postural changes inherent to motor development from sitting/crawling at 9 months of age to ease of walking/standing by 18 months may also affect acoustics in the child's vocal production. These developmental processes change the baseline acoustic properties of the

child's vocalizations (Scheiner et al., 2002; Warlaumont et al., 2010), thereby impacting the overall relationship between acoustic features of vocal expressions and the child's emotional state.

In more specific behavioral contexts, such as Crying, Babble, and Other Emoting, however, we have observed notable longitudinal progressions that point to changes in correlation of acoustic vocal features with child emotion over time. In Crying, there is an increasing trend, which we see most clearly for ***child*** models. Although the ***combined*** models also seem to be increasing in adjusted R-squared, there is no consistent trend in the adult-only models to suggest any adult component to this progression in the combined models. Babble is characterized by clear downward trends, most strikingly for ***adult surrounding*** models, and also for ***child*** and ***combined*** models. Other Emoting bears an upward trend in both ***child*** and ***combined*** models, but no such pattern in adult-only models.

What kind of developmental phenomena might be responsible for such progressive changes in correlation as we have observed in Crying, Babble, and Other Emoting? Towards answering this question, we propose several hypotheses that begin to portray a connection, albeit loose, between the three contexts' longitudinal progressions. Our reasoning takes into account both physiological and cognitive development, as follows:

Babble is a process of experimentation and play towards learning how to control the developing vocal tract to produce speech (Petitto & Marentette, 1991; Vihman et al., 1985; Warlaumont et al., 2010). This mastery develops gradually, ultimately evolving into speech, and in the process opening a new medium for general-purpose communication. In early stages of babble, the child's lack of intentional control of the vocal tract for speech purposes may more instinctively and reflexively express the child's emotional state as a consequence of subconscious physiological responses originating in the autonomic nervous system (Fox & Davidson, 1986; Oudeyer, 2003; Scherer, 2003; Sundberg, 1977; Ververidis & Kotropoulos, 2006). As the child gains more control of the vocal tract, conscious intention in channeling Babble for general-purpose communication begins to take over, and Babble loses this intimate physiological connection to affect. This new medium for communicative intent also offers the child new options for expressing complex emotions that just cannot be conveyed by crying or simple emoting. The variety and nuanced subtlety of these new

emotions may make them harder to distinguish acoustically, not only for our model, leading to decreasing adjusted R-squared, but also for the human caretaker, who would therefore be much less likely to elicit consistent, emotion-specific caretaker responses (Zeifman, 2005; Zeskind, 2005). The latter may contribute to the striking decrease in correlation that we observed between the child's emotional state and acoustic features of surrounding adult speech. Further, Bloom found that learning words and expressing emotion compete for the child's limited cognitive resources, observing that neutral affect promotes language development during the 9-17 month age period (Bloom, 1998). The developmental transition from Babble to Speech therefore seems to be related to the child's ability to maintain a neutral emotional state (Bloom, 1998). Greater emotional self-control brings with it less of a need for extrinsic emotion regulation by caretakers (Cole et al., 2004; Eisenberg & Spinrad, 2004; Trevarthen, 1993), which might also explain the dramatic downward trend that we observe for ***adult surrounding*** models.

Along the same lines, the upward trends in correlation between the child's emotion and the child's vocal expression that we see in Crying and Other Emoting seem to reflect an inverse relationship to the downward trends in Babble. In early stages of infancy, Crying and Other Emoting are a child's primary means of communication, with a rather varied repertoire of cry types and sounds (Zeifman, 2005; Zeskind, 2005). As Babble and Speech progressively take the place of Crying and Other Emoting as the child's general-purpose mode of communication, Crying and Other Emoting gradually transition to a mode that is reserved for expressing only raw emotion. This specialization may serve to reduce the variability in the repertoire of acoustic patterns while increasing the consistency of emotional expression. A longitudinal study by Scheiner et al (2002) of infants during the first year of life supports this hypothesis, having found increasing homogeneity of acoustic properties of child cries with age, including a decrease in frequency range. Development of emotion differentiation may also be a factor in these upward trends (Scheiner et al., 2002; Trevarthen, 1993):

> There is no general agreement, however, whether infants during their first months of life are able to express specific emotions in their behavior at all (for an overview, see Strongman (Strongman, 1996)). Some authors suggest that, in the first months, the expressive behavior is to a large extent random.
>
> (Scheiner et al., 2002)

> The evidence is clear that infants possess at birth, not only a coherent and differentiated emotional system…but also the distinctions between 'person-related', 'thing-related', and 'body-related' functions of emotions. **Admittedly, these functions become clearer and more effective with development**, but they appear in rudimentary form in the newborn.
>
> (Trevarthen, 1993)

Thus, with growing specificity of emotions and mastery of vocal tract control, the child may be increasingly able to express subtle nuances that more accurately reflect his emotional state, resulting in increased correlation between the child's emotional state and vocal expression.

6.3 Dyadic Considerations

We included adult-only models in our analysis to explore the degree to which caretaker speech reflects the emotional state of the child. Despite the fact that adult-only models generally have much smaller adjusted R-squared than ***child*** or ***combined*** models, they seem to reveal some notable insights about the dyadic context surrounding a child's emotional state. This significantly lower correlation of adult-only models is to be expected, since it is natural for a child's vocalizations to correspond most closely to the child's own emotional state. Even so, adult-only models achieve medium effect sizes in many contexts, including the overall dataset, which means that there are some fairly significant correlations between adult speech and perceived child emotion, even when looking at the dataset as a whole.

Within more specific contexts, we find multiple patterns that seem to derive plausible explanations from typical behavioral tendencies. In Social Crying, for example, it seems natural for adults to express a characteristic repertoire of soothing tones in comforting the child, which may explain the peak in adult-only adjusted R-squared for this context, as noted in Section 6.1. Likewise, the relatively high correlation of adult speech ***after*** a child's Social Laughing vocalization with the child's emotional state may reflect a pattern of caretakers laughing in response to the child's laughter, which would seem to be a natural, common occurrence in a social context. The lower overall correlation of adult speech with child emotion in Babble and Speech contexts, becoming negligible for Social Speech, also

makes sense because Babble and Speech carry a separate primary purpose beyond purely communicating emotion, as discussed in Section 6.2. With the exception of Social Speech, however, adult-only models consistently demonstrate higher correlations in Social Situations for each behavioral context, which seems to reflect the fact that adult speech more purposefully addresses the child's overall emotional state in social situations. In addition to providing insights into behavioral patterns within the child-caretaker dyad, such observations in our adult-only results that mirror well-known caretaker tendencies serve to further substantiate our methodology.

In exploring caretaker-specific models, our dyadic analysis also highlights individual differences among caretakers, such as the Nanny correlating best in Social Crying contexts, the Mother far exceeding the other caretakers in Laughing contexts, and the Father most being highly correlated with the child's emotional state in Social Babble and Social Speech. Most notably, the Nanny's speech tends to be significantly more correlated with the child's emotional state, overall, in most contexts. While such insights could help optimize perceptual accuracy by taking into account the identity of the caretaker, we find that models built using all caretakers in aggregate are generally more successful than each caretaker individually.

The caretaker-specific trends, however, begin to draw out a tapestry of family dynamics that could be useful in applications other than simply determining the child perceived emotional state. One interpretation of adjusted R-squared in adult-only models is as a measure of intersubjectivity, or affect synchrony. We therefore envision the methodology of our dyadic analysis to be of benefit in a variety of developmental research areas, such as:

- Predicting socio-emotional outcomes using mother-infant relationship markers (Deater-Deckard & Petrill, 2004; Forcada-Guex et al., 2006; Symons, 2001)

- Studying the relationship between extrinsic emotion regulation and the development of self-control in child behavior (Cassidy, 1994; Cole et al., 2004; Eisenberg & Spinrad, 2004; Feldman et al., 1999)

- Investigating caretaker-specific factors affecting child language acquisition (Bloom, 1998; Vosoughi et al., 2010)

- Studying developmental pathologies in affect synchrony, such as those seen in attachment disorders and autism (Greenspan, 2001; Schore, 2001)

6.4 Concluding Thoughts

In this work, we have demonstrated that a high-performing model for automatically perceiving a child's emotional state from vocal acoustic features can be created within the Speechome corpus for a single, typically developing child. We have developed a methodology for modeling acoustic correlates of the child's perceived emotional state within the Speechome corpus, and we implement this methodology through a process of manual annotation, data mining, acoustic feature extraction, and Partial Least Squares regression. Comparative analysis of perceptual model performance across longitudinal and socio-behavior contexts has yielded some notable developmental and dyadic insights.

The manual annotation step remains a necessary bottleneck in our methodology, in order to establish the ground truth about the child's observed emotional state using human perception. We have devised and applied several data mining strategies as part of our methodology for the purpose of minimizing annotation time, including the use of Speaker ID metadata to filter out portions of the audio that are not likely to contain child vocalizations, which reduced annotation time by a total of 87 hours. Further, the results of our exploratory analysis suggest a possibility for simplifying our annotation methodology in future applications. We observe a consistency across socio-behavioral contexts, implying that there is not much to gain by asking annotators to indicate the behavioral nature of a vocalization, or whether it occurs within a social situation. However, we recommend that the socio-behavioral component of our methodology be retained, clarified to increase inter-annotator agreement, and tested in Speechome datasets of several other children once they become available, in order to test the generalizability of this pattern.

In modeling acoustic correlates of the child's perceived emotional state, we have found Partial Least Squares regression to be an effective approach. PLS regression has enabled us to capture, with robustness against overfitting, the relationship between a highly multivariate set of input variables, consisting of 82 acoustic vocal features, and a multivariate outcome, namely the child's emotional state as represented by Schlosberg's two-dimensional space for characterizing affect. Using adjusted R-squared to evaluate our models' perceptual accuracy has provided us with a versatile metric with many synonymous interpretations, including goodness of fit, which implies correlation. In evaluating adult-only models, adjusted R-squared may be a functional metric for intersubjectivity. Through the use of cross validation, we include generalizability considerations in our evaluation of perceptual accuracy; however, we recommend further investigation with larger sample sizes and explicit training/testing datasets to confirm our results.

In addition to evaluating perceptual accuracy, our exploratory analysis has also brought forward several longitudinal insights, which point to developmental shifts that may occur as the child gains mastery of the vocal tract and increases in cognitive abilities and awareness. Because monthly models tend to outperform time-aggregate models while still remaining consistent over time in the general case, we deduce that baseline properties of a child's vocal emotional expression (i.e. the most significant vocal acoustic correlates) may change over time, without changing the correlation itself. In relation to emotional expressiveness and intersubjectivity, our exploratory observations of increasing and decreasing trends in more specific contexts lead us to hypothesize an inverse relationship between a general-purpose communication medium such as babble and pure emoting vocalizations such as crying.

We qualify all of our findings with a caveat: the high perceptual accuracy, longitudinal progressions, dyadic patterns, and consistencies across contexts that we have observed represent the vocal emotional expression of a single, typically developing child. Our methodology might reveal very different results in a child who may be more inhibited temperamentally and therefore less emotionally expressive. In addition to temperament, there are many factors that can lead to differences in emotional and language development among typically developing children, such as innate attachment style, parental

responsiveness, and gender. We also expect results to diverge in the presence of developmental disorders such as autism, in which emotional expression and affect synchrony are impaired (Greenspan, 2001; Macdonald, 2006).

6.4.1 Future directions

As we note above, the Speechome corpus currently represents data for only a single child. To fully harness the exciting new possibilities for the study of child development using the kind of dense, longitudinal, ecologically-valid data that the Human Speechome Project has pioneered, similar datasets need to be recorded and assembled across multiple children. This diversity would enable tests to evaluate the generalizability of methodologies and insights developed around the Speechome dataset, as well as comparative analyses between individuals within a particular group (e.g. age, gender) and between subgroups, such as neurotypical and developmentally impaired children, in the study of developmental disorders.

A major obstacle to such large-scale deployment has been the cost and complexity involved in the original recording installation, where cameras and microphones were installed in every room of the house, and great care and much expense were taken to integrate the system seamlessly into the home (D. Roy, 2009). To address this, a compact, portable version of the embedded Speechome recording system has been developed, called the Speechome Recorder. The first prototype of the Speechome Recorder is shown in Figure 6-1.

In addition to a similar overhead microphone and camera placement, the Speechome recorder introduces a second, frontal camera. This frontal view has been added to facilitate analysis of facial expressions and gestures, which are elusive from the bird's eye view of the overhead camera. Future research in characterizing emotion within the Speechome corpus using objective measures may benefit from multimodal models that integrate facial expressions, gestures, posture (Begeer et al., 2006; Busso et al., 2004; El Kaliouby et al., 2006), and even physiological sensor metrics (Fox & Davidson, 1986; Goodwin et al., 2006; Liu et al., 2008; Picard et al., 2001), with vocal acoustic correlates.

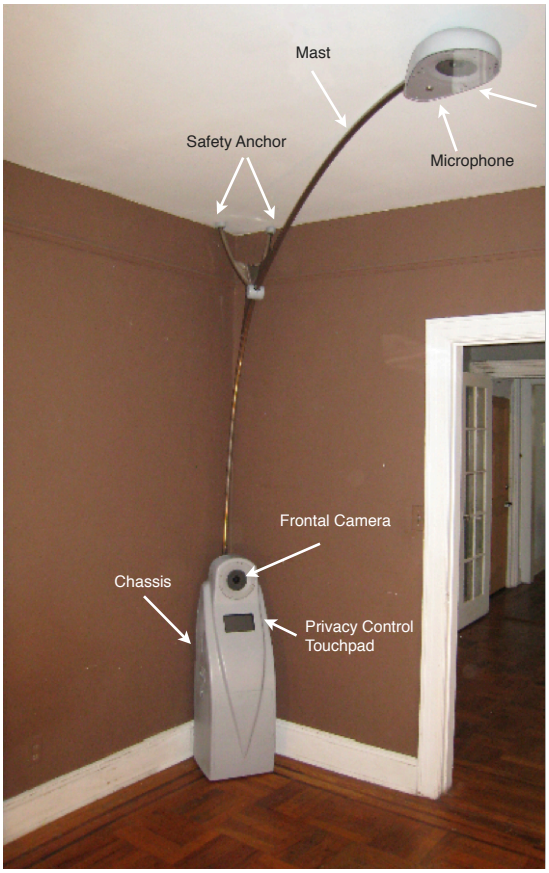


Figure 6-1. Speechome Recorder

In the broader view, for future application, our methodology intends to create a separate perceptual model for every child, tailored specifically to each child's temperament, idiosyncrasies of vocal expression, developmental age, and caretakers. As new Speechome datasets for multiple children are collected using the Speechome Recorder, it is our hope that this methodology, applied to each child's dataset, will not only automate the extraction of each child's affective state, but also yield comparative insights across children about how emotional expression and caretaker intersubjectivity varies from child to child, and between neurotypical and developmentally impaired children.

Through such large-scale comparative studies, deployment of the Speechome Recorder may facilitate the study of developmental milestones that can potentially reveal insights into how autism and other developmental disorders evolve from infancy through early childhood. In addition, we envision the possible application of Speechome technologies towards implementing non-invasive continuous monitoring devices that can

inform parents of subtle deviations in their child's behavior and development. For instance, due to the wide variability in the autistic phenotype, clinicians and caretakers in this area have acknowledged a need for continuous monitoring and surveillance technologies after diagnosis to better inform the choice of treatments and accommodations that would best serve a particular individual (Hayes et al., 2008; Hayes et al., 2004; Kientz et al., 2005; Klin et al., 2005; Lord, 2000; Zwaigenbaum et al., 2009). Further, an effective course of treatment requires continued monitoring after intervention to validate its efficacy and quantify the child's response to treatment (Lord et al., 2005).

Automated mechanisms and methodologies designed to support empirical study and analysis of Speechome's dense, longitudinal, ecologically valid observational data hold great promise for quantifying early processes of atypical development that may lead to earlier detection and more individualized treatment. A better understanding of atypical processes may also provide new insights into neurotypical development. In light of the significance of emotional factors in research pertaining to child development, we hope that the work of this thesis has brought Speechome one step closer to facilitating such advances.

Bibliography

Abdi, H. (2003) Partial Least Squares (PLS) Regression. In M. Lewis-Beck, A. Bryman & T. Futing (Eds.), *Encyclopedia of Social Sciences Research Methods*. Thousand Oaks, CA: Sage.

Adolph, K., Robinson, S. R., Young, J. W., & Gill-Alvarez, F. (2008) What is the shape of developmental change? *Psychological Review 115*(3), 527-543.

Ali, A., Bhatti, S., & Mian, M. S. (2006) Formants Based Analysis for Speech Recognition, *IEEE International Conference on Engineering of Intelligent Systems*.

Altman, D. G. (1991) *Practical Statistics for Medical Research*: Chapman & Hall.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002) Prosody-based automatic detection of annoyance and frustration in human-computer dialog, *Proc. Int. Conf. Spoken Language Processing (ICSLP'02)* (Vol. 3, pp. 2037-2040).

Bachorowski, J., Smoski, M. J., & Owren, M. J. (2001) The acoustic features of human laughter. *The Journal of the Acoustical Society of America 110*, 1581-1597.

Bachorowski, J. A., & Owren, M. J. (1995) Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychol Sci 6*(4), 219-224.

Banse, R., & Scherer, K. R. (1996) Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology 70*, 614-636.

Baranek, G. (1999) Autism during infancy: A retrospective video analysis of sensory-motor and social behaviors at 9-12 months of age. *Journal of Autism and Developmental Disorders 29*(3), 213-223.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., et al. (2006) Combining Efforts for Improving Automatic Classification of Emotional User States, *Proc. IS-LTC* (pp. 240-245). Ljubliana.

Begeer, S., Rieffe, C., Meerum Terwogt, M., & Stockmann, L. (2006) Attention to facial emotion expressions in children with autism. *Autism 11*, 503-521.

Bell, M. A., & Wolfe, C. D. (2004) Emotion and Cognition: An Intricately Bound Developmental Process. *Child Development 75*(2).

Beritelli, F., Casale, S., Russo, A., & Serrano, S. (2006) Speech emotion recognition using MFCCs extracted from a mobile terminal based on ETSI front end, *International Conference on Signal Processing*.

Bloom, L. (1973) *One word at a time*. The Hague: Mouton.

Bloom, L. (1998) Language Development and Emotional Expression. *Pediatrics 102*(5), 1272-1277.

Boersma, P. (1993) *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*: Institute of Phonetic Sciences, University of Amsterdam.

Boersma, P., & Weenink, D. (2010) Praat: doing phonetics by computer (Version 5.1.37).

Bolnick, R., Spinrad, T., Eisenberg, N., Kupfer, A., & Liew, J. (2006) Predicting language development from early emotional expressivity, *Poster session presented at the Biennial Meeting of the International Society on Infant Studies*. Kyoto, Japan.

Bowlby, J. (1973) *Attachment and loss: Vol 2. Separation*. New York: Basic Books.

Breazeal, C. (2001) Emotive qualities in robot speech, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2001)* (pp. 1389-1394).

Breazeal, C., & Aryananda, L. (2002) Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots 12*(1), 83-104.

Bruner, J. (1983) *Child's Talk: Learning to Use Language*: Norton.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005) A database of German emotional speech, *Proc. Interspeech* (pp. 1517-1520).

Burkhardt, F., & Sendlmeier, W. (2000) Verification of acoustical correlates of emotional speech using formant-synthesis, *Proceedings of the ISCA Workshop on Speech and Emotion*.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., et al. (2004) Analysis of Emotion Recognition using Facial Expressions, Speech, and Multimodal Information, *ICMI'04*. State College, PA.

Busso, C., & Narayanan, S. S. (2008) The expression and perception of emotions: Comparing assessments of self versus others, *Interspeech* (pp. 257-260). Brisbane, Australia.

Calkins, S. D., & Bell, M. A. (2009) *Child Development at the Intersection of Emotion and Cognition*: American Psychological Association.

Camacho, A., & Akiskal, H. S. (2005) Proposal for a bipolar-stimulant spectrum: temperament, diagnostic validation, and therapeutic outcomes with mood stabilizers. *Journal of Affective Disorders 85*, 217-230.

Campbell, N. (2001) Building a corpus of natural speech -- and tools for the processing of expressive speech -- the JST CREST ESP project, *In Proceedings of the 7th European Conference on Speech Communication and Technology* (pp. 1525-1528).

Cassidy, J. (1994) Emotion Regulation: Influences of Attachment Relationships. *Monographs of the Society for Research in Child Development 59*(2-3), 228-249.

Childers, D. G. (1978) *Modern spectrum analysis*. New York: IEEE Press.

Chuang, Z. J., & Wu, C. H. (2004) Multi-Modal Emotion Recognition from Speech and Text. *Computational Linguistics and Chinese Language Processing 9*(2), 45-62.

Clement, C. J., Koopmans-van Beinum, F. J., & Pols, L. C. W. (1996) Acoustical characteristics of sound production of deaf and normally hearing infants, *Fourth International Conference on Spoken Language Processing* (Vol. 3, pp. 1549-1552). Philadelphia, PA.

Clemins, P. J., & Johnson, M. T. (2006) Generalized perceptual linear prediction features for animal vocalization analysis. *Journal of the Acoustical Society of America 120*(1), 527-534.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Edu. and Psych. Meas 20*, 37-46.

Cohen, J. (1992) A power primer. *Psychological Bulletin 112*, 155-159.

Cole, P. M., Martin, S. E., & Dennis, T. A. (2004) Emotion regulation as a scientific construct: Methodological challenges and directions for child development research. *Child Development 75*, 317-333.

Cole, P. M., Michel, M. K., & Teti, L. O. D. (1994) The development of emotion regulation and dysregulation: A clinical perspective. *Monographs of the Society for Research in Child Development 59*(2/3).

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. S. (2001) *Introduction to Algorithms* (Second Edition ed.): MIT Press.

Dautenhahn, K. (1999) Embodiment and interaction in social intelligent life-like agents. In *Computation for Metaphors, Analogy, and Agents* (Vol. 1562/1999, pp. 102-141). Heidelberg: Springer Berlin.

Davis, S. B., & Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, Signal Processing 28*(4), 357-366.

Dawson, G. (1991) A psychobiological perspective on the early socioemotional development of children with autism. In S. Toth & D. Cichetti (Eds.), *Rochester symposium on developmental psychology* (Vol. 3, pp. 207-234). Hilldale, NJ: Lawrence Erlbaum.

De Giacomo, A., & Fombonne, E. (1998) Parental recognition of developmental abnormalities in autism. *European Child and Adolescent Psychiatry 7*, 131-136.

De Vries, S., & Ter Braak, C. J. F. (1995) Prediction error in partial least squares regression: a critique on the deviation used in The Unscrambler. *Chemometrics and Intelligent Laboratory Systems 30*(2), 239-245.

Deater-Deckard, K., & Petrill, S. A. (2004) Parent-child dyadic mutuality and child behavior problems: an investigation of gene-environment processes. *Journal of Child Psychology and Psychiatry 45*(6), 1171-1179.

Dellaert, F., Polzin, T., & Waibel, A. (1996) Recognizing emotion in speech, *International Conference on Spoken Language Processing*. Philadelphia, PA.

Dominey, P., & Dodane, C. (2004) Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Journal of Neurolinguistics 17*, 121-145.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003) Emotional speech: Towards a new generation of databases. *Speech Communication 40*, 33-60.

Douglas-Cowie, E., Cowie, R., & Schroder, M. (2000) A new emotion database: considerations, sources, and scope, *Proc. ISCA Workshop on Speech and Emotion* (pp. 39-44).

Eisenberg, N., & Spinrad, T. L. (2004) Emotion-related Regulation: Sharpening the Definition. *Child Development 75*(2), 334-339.

El Kaliouby, R., Picard, R. W., & Baron-Cohen, S. (2006) Affective computing and autism. *Annals of the New York Academy of Sciences 1093*(1), 228-248.

Feldman, R., Greenbaum, C. W., & Yirmiya, N. (1999) Mother-infant affect synchrony as an antecedent of the emergence of self-control. *Developmental Psychology 35*(5), 223-231.

Feldman-Barrett, L. A. (1995) Variations in the Circumplex Structure of Mood. *Personality and Social Psychology Bulletin 21*, 806-817.

Fischer, K. (1999) *Annotating emotional language data*: Univ. of Hamburg.

Fischer, K. W., Shaver, P. R., & Carnochan, P. (1990) How emotions develop and how they organise development. *Cognition & Emotion*.

Forcada-Guex, M., Pierrehumbert, B., Borghini, A., Moessinger, A., & C., M.-N. (2006) Early dyadic patterns of mother-infant interactions and outcomes of prematurity at 18 months. *Pediatrics 118*(1), e107-114.

Fox, N. A., & Davidson, R. J. (1986) Psychophysiological measures of emotion: new directions in developmental research. In C. E. Izard & P. B. Read (Eds.), *Measuring Emotions in Infants and Children* (Vol. II, pp. 13-50). New York, NY: The Press Syndicate of the University of Cambridge.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000) Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. Biomedical Engineering 7*, 829-837.

Fuller, B. (1991) Acoustic discrimination of three types of infant cries. *Nursing Research 40*(3), 336-340.

Garcia, J. O., & Garcia, C. A. R. (2003) Mel-Frequency Cepstrum Coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks, *Proceedings of the International Joint Conference on Neural Networks* (pp. 3140-3145).

Garthwaite, P. H. (1994) An interpretation of Partial Least Squares. *Journal of the American Statistical Association 89*(425), 122-127.

Goodwin, M. S., Groden, J., Velicer, W. F., Lipsitt, L. P., Baron, M. G., Hofmann, S. G., et al. (2006) Cardiovascular arousal in individuals with autism. *Focus on Autism and Other Developmental Disabilities 21*(2), 100-123.

Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S., et al. (1995) Representation of prosodic and emotional features in a spoen language database, *Proc. of XIIIth ICPhS* (Vol. 1, pp. 242-245). Stockholm.

Greasley, P., Sherrard, C., & Waterman, M. (2000) Emotion in language and speech: Methodological issues in naturalistic approaches. *Language and Speech 43*, 355-375.

Greasley, P., Sherrard, C., Waterman, M., Setter, J., Roach, P., Arnfield, S., et al. (1996) The perception of emotion in speech. *Abs Int J Psychol 31*(3/4), 406.

Greenspan, S. I. (2001) The affect diathesis hypothesis: The role of emotions in the core deficit in autism and in the development of intelligence and social skills. *Journal of Developmental and Learning Disorders 5*(1).

Gustafson, G. E., & Green, J. A. (1989) On the importance of fundamental frequency and other acoustic features in cry perception and infant development. *Child Development 60*, 772-780.

Hailpern, J., & Hagedorn, J. (2008) VCode and VData: Illustrating a new Framework for Supporting the Video Annotation Workflow, *AVI '08*. Naples, Italy: ACM.

Hansen, J. H. L., & Cairns, D. A. (1995) ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments. *Speech Communication 16*, 391-422.

Harlow, L. L. (2005) *The Essence of Multivariate Thinking*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hart, J. t., Collier, R., & Cohen, A. (1990) *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*: Cambridge University Press.

Hayes, G. R., Gardere, L. M., Abowd, G. D., & Truong, K. N. (2008) CareLog: A selective archiving tool for behavior management in schools, *CHI 2008 Proceedings - Tools for Education*. Florence, Italy.

Hayes, G. R., Kientz, J. A., Truong, K. N., White, D. R., Abowd, G. D., & Pering, T. (2004) Designing capture applications to support the education of children with autism, *Ubiquitous Computing*. Nottingham, UK.

Hoskuldsson, P. (1988) PLS regression methods. *Journal of Chemometrics 2*, 211-228.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., & Nogueiras, A. (2002) Interface databases, design and collection of a multilingual emotional speech database, *Proceeding of the 3rd Language Resources and Evaluation Conference* (pp. 2024-2028).

Hubert, M., & Vanden Branden, K. (2003) Robust methods for Partial Least Squares regression. *Journal of Chemometrics 17*, 537-549.

Hudenko, W. J., Stone, W., & Bachorowski, J. (2009) Laughter differs in children with autism: an acoustic analysis of laughs produced by children with and without the disorder. *Journal of Autism and Developmental Disorders 39*(10), 1392-1400.

Jensen, J. H., Christensen, M. G., Ellis, D. P. W., & Jensen, S. H. (2009) Quantitative Analysis of a Common Audio Similarity Measure. *IEEE Transactions on Audio, Speech, and Language Processing 17*(4), 693-703.

Johnson, M. L., Veldhuis, J. D., & Lampl, M. (1996) Is growth saltatory? The usefulness and limitations of frequency distributions in analyzing pulsatile data. *Endocrinology 137*, 5197-5204.

Kagan, J. (1994) On the nature of emotion. *Monographs of the Society for Research in Child Development 59*(2/3), 7-24.

Kalivas, J. H. (1997) Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems 37*, 255-259.

Kamaruddin, N., & Wahab, A. (2008) Speech emotion verification system (SEVS) based on MFCC for real time applications, *4th International Conference on Intelligent Environments*. Seattle, WA.

Keller, G. (2009) *Statistics for Management and Economics* (8th Edition ed.). Mason, OH: South-Western Cengage Learning.

Kent, R. D., & Murray, A. D. (1982) Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *Journal of the Acoustical Society of America 72*(2), 353-365.

Kientz, J. A., Boring, S., Abowd, G. D., & Hayes, G. R. (2005) Abaris: Evaluating automated capture applied to structured autism interventions, *Ubiquitous Computing* (Vol. 3660, pp. 323-339).

Klasmeyer, G., & Sendlmeier, W. F. (1995) Objective voice parameters to characterize the emotional content in speech, *13th International Congress of Phonetic Sciences*. Stockholm.

Klin, A., Saulnier, C., Tsatsanis, K., & Volkmar, F. R. (2005) Clinical evaluation in autism spectrum disorders: psychological assessment within a transdisciplinary framework. In F. R. Volkmar, R. Paul, A. Klin & D. Cohen (Eds.), *Handbook of Autism and Pervasive Developmental Disorders* (pp. 772-798). Hoboken, NJ: John Wiley and Sons.

Klinnert, M. D., Campos, J. J., Sorce, J. F., Emde, R. N., & Svejda, M. (1983) Emotions as behavior regulators: Social referencing in infancy. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience* (pp. 57-86). New York: Academic Press.

Kovacic, G., & Boersma, P. (2006) Spectral characteristics of three styles of Croatian folk singing. *Journal of the Acoustical Society of America 119*(3), 1805-1816.

Kovacic, G., Boersma, P., & Domitrovic, H. (2003) Long-term average spectra in professional folk singing voices: A comparison of the Klapa and Dozivacki styles, *Proceedings of the Institute of Phonetic Sciences* (pp. 53-64): University of Amsterdam.

Lampl, M., Johnson, M. L., & Frongillo, E. A. (2001) Mixed distribution analysis identifies saltation and stasis growth. *Annals of Human Biology 28*(403-411).

Lay New, T., Wei Foo, S., & De Silva, L. (2003) Speech Emotion Recognition Using Hidden Markov Models. *Speech Communication 41*(4), 603-623.

Lee, C. M., & Narayanan, S. S. (2005) Towards detecting emotions in spoken dialogs. *IEEE Trans. Speech and Audio Process. 13*(2), 293-303.

Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., et al. (2004) Emotion Recognition based on Phoneme Classes, *International Conference on Spoken Language Processing*. Jeju Island, Korea.

Lemerise, E. A., & Arsenio, W. F. (2000) An integrated model of emotion processes and congition in social information processing. *Child Development 71*(1), 107-118.

Liscombe, J., Riccardi, G., & Hakkani-Tur, D. (2005) Using Context to Improve Emotion Detection in Spoken Dialog Systems. *Proceedings of Interspeech*, 1845-1848.

Liu, C., Conn, K., Sarkar, N., & Stone, W. (2008) Physiology-based affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder. *Int'l Journal of Human-Computer Studies 66*, 662-677.

Lord, C. (2000) Commentary: Achievements and future directions for intervention research in communication and Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders 30*(5), 393-398.

Lord, C., Wagner, A., Rogers, S., Szatmari, P., Aman, M., Charman, T., et al. (2005) Challenges in evaluating psychosocial interventions for Autistic Spectrum Disorders. *Journal of Autism and Developmental Disorders 35*(6), 695-708.

Macdonald, H. (2006) Recognition and expression of emotional cues by autistic and normal adults. *Journal of Child Psychology and Psychiatry 30*(6), 865-877.

Martland, P., Whiteside, S. P., Beet, S. W., & Baghai-Ravary, L. (1996) Estimating child and adolescent formant frequency values from adult data, *Proceedings of the Applied Science and Engineering Laboratories Conference ICSLP'96* (pp. 622-625).

McClellan, J. H., Schafer, R. W., & Yoder, M. A. (1999) *DSP First: A Multimedia Approach*: Prentice-Hall, Inc.

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996) Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage 3*, 143-157.

McIntosh, A. R., Chau, W. K., & Protzner, A. B. (2004) Spatiotemporal analysis of event-related fMRI data using partial least squares. *Neuroimage 23*, 764-775.

Miller, M. (2009) Speaker Identification in the Human Speechome Corpus using Explicit Noise Modeling: Massachusetts Institute of Technology.

Molnar, C., Kaplan, F., Roy, P., Pachet, F., Pongracz, P., Doka, A., et al. (2008) Classification of dog barks: a machine learning approach. *Animal Cognition 11*(3), 389-400.

Morgan, K. (2006) Is autism a stress disorder? In *Stress and Coping in Autism*.

Mozziconacci, S. J. L., & Hermes, D. J. (2000) Expression of emotion and attitude through temporal speech variations, *Int'l Conference on Spoken Language Processing* (Vol. 2, pp. 373-378). Beijing.

Murray, I. R., & Arnott, J. L. (1993) Toward the simulation of emotion in synthetic speech: a review of the literature on vocal emotion. *Journal of the Acoustical Society of America 93*(2), 1097-1108.

Navas, E., Castelruiz, A., Luengo, I., Sanchez, J., & Hernaez, I. (2004) Designing and recording an audiovisual database of emotional speech in basque, *Proc. LREC*.

Neiberg, D., Elenius, K., & Laskowski, K. (2006) Emotion recognition in spontaneous speech using GMMs, *International Conference on Spoken Language Processing* (pp. 809-812).

Nelson, K. (1974) Concept, word, and sentence: Interrelations in acquisition and development. *Psychological Review 81*, 267-285.

Noonan, K. J., Farnum, C. E., Leiferman, E. M., Lampl, M., Markel, M. D., & Wilsman, N. J. (2004) Growing pains: Are they due to increased growth during recombency as documented in a lamb model? *Journal of Pediatric Orthopedics 24*, 726-731.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003) Speech emotion recognition using hidden Markov models. *Elsevier Speech Communications Journal 41*(4), 603-623.

Nwokah, E., Hsu, H., & Fogel, A. (1993) Vocal affect in three-year-olds: A quantitative acoustic analysis of child laughter. *The Journal of the Acoustical Society of America 94*, 3076-3090.

Ochs, E., & Schieffelin, B. (1989) Language has a heart. *Text 9*, 7-25.

Ortony, A., Clore, G., & Collins, A. (1988) Vocal expression and perception of emotion. *Current Direct Psychol Sci 8*(2), 53-57.

Osgood, C., Suci, G., & Tannenbaum, P. (1957) *The Measurement of Meaning*. Urbana: University of Illinois Press.

Oudeyer, P.-Y. (2003) The production and recognition of emotional speech: features and algorithms. *Int'l Journal of Human-Computer Studies 59*, 157-183.

Papaeliou, C., Minadakis, G., & Cavouras, D. (2002) Acoustic patterns of infant vocalizations expressing emotions and communicative functions. *Journal of Speech Language and Hearing Research 45*, 311-317.

Petitto, L. A., & Marentette, P. F. (1991) Babbling in the manual mode: evidence for the ontogeny of language. *Science 251*(5000), 1493-1496.

Petroni, M., Malowany, A. S., Johnston, C. C., & Stevens, B. J. (1994) A New, Robust Vocal Fundamental Frequency (F0) Determination Method for the Analysis of Infant Cries., *IEEE Seventh Symposium on Computer-Based Medical Systems* (pp. 223-228).

Petroni, M., Malowany, A. S., Johnston, C. C., & Stevens, B. J. (1995) Identification of pain from infant cry vocalizations using artificial neural networks (ANNs), *Proc. of SPIE* (Vol. 2492, pp. 729-738).

Petroni, M., Malowany, A. S., Johnston, C. C., & Stevents, B. J. (1994) A Crosscorrelation-Based Method for Improved Visualization of Infant Cry Vocalizations, *Canadian Conference on Electrical and Computer Engineering* (pp. 453-456).

Petrovich-Bartell, N., Cowan, N., & Morse, P. A. (1982) Mothers' Perceptions of Infant Distress Vocalizations. *Journal of Speech and Hearing Research 25*, 371-376.

Petrushin, V. A. (1999) Emotion in speech recognition and application to call centers, *Proc. Artificial Neural Networks in Engineering (ANNIE 99)* (Vol. 1, pp. 7-10).

Picard, R. W., Vyzas, E., & Healey, J. (2001) Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*(10), 1175-1191.

Pirouz, D. M. (2006) An overview of partial least squares: University of California, Irvine.

Rahurkar, M., & Hansen, J. H. L. (2002) Frequency band analysis for stress detection using a Teager energy operator based feature., *Proc Int Conf Spoken Language Processing (ICSLP'02)* (Vol. 3, pp. 2021-2024).

Rank, E., & Pirker, H. (1998) Generating emotional speech with a concatenative synthesizer, *Fifth International Conference on Spoken Language Processing* (Vol. 3, pp. 671-674). Sydney.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S., & Setter, J. (1998) Transcription of prosodic and paralinguistic features of emotional speech. *J Int Phonetic Assoc 28*, 83-94.

Rosipal, R., & Trejo, L. J. (2001) Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research 2*, 97-123.

Rothbart, M. K. (2005) Early temperament and psychosocial development. In R. E. Tremblay, R. G. Barr & R. D. e. V. Peters (Eds.), *Encyclopedia on Early Childhood Development* (pp. 1-6). Montreal, Quebec: Centre of Excellence for Early Childhood Development.

Rothbart, M. K., & Posner, M. I. (1985) Temperament and the development of self-regulation. In H. Hartledge & C. R. Telzrow (Eds.), *Neuropsychology of Individual Differences*. New York: Plenum.

Roy, B. C. (2007) *Human-machine collaboration for rapid speech transcription.* Massachusetts Institute of Technology, Cambridge, MA.

Roy, B. C., & Roy, D. (2009) Fast transcription of unstructured audio recordings., *Proceedings of Interspeech*. Brighton, England.

Roy, D. (2009) New Horizons in the Study of Child Language Acquisition, *Proceedings of Interspeech*. Brighton, England.

Roy, D., Patel, P., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., et al. (2006) The Human Speechome Project, *28th Annual Conference of the Cognitive Science Society*.

Ruvolo, P., & Movellan, J. (2008) Automatic cry detection in early childhood education settings, *7th IEEE International Conference on Development and Learning* (Vol. 7, pp. 204-208).

Sanson, A., Letcher, P., Smart, D., Prior, M., Toumbourou, J. W., & Oberklaid, F. (2009) *Associations between early childhood temperament clusters and later psychosocial adjustment*.

Scassellati, B. (2005) How social robots will help us to diagnose, treat, and understand autism, *Proceedings of the 12th international symposium of robotics research (ISSR'05)*. San Francisco, CA.

Scheiner, E., Hammerschmidt, K., Jurgens, U., & Zwirner, P. (2002) Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants. *Journal of Voice 16*(4), 509-529.

Scherer, K. R. (2003) Vocal communication of emotion: A review of research paradigms. *Speech Communication 40*, 227-256.

Schiel, F., Steininger, S., & Turk, U. (2002) The Smartkom multimodal corpus at BAS, *Proc. Language Resources and Evaluation (LREC'02)*.

Schlosberg, H. (1952) The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology 44*, 229-237.

Schonweiler, R., Kaese, S., Moller, S., Rinscheid, A., & Ptok, M. (1996) Neuronal networks and self-organizing maps: new computer techniques in the acoustic evaluation of the infant cry. *Int'l J Pediatr Otorhinolaryingol 38*, 1-11.

Schore, A. N. (2001) Effects of a secure attachment relationship on right brain development, affect regulation, and infant mental health. *Infant Mental Health Journal 22*(1-2), 7-66.

Schroder, M. (2001) Emotional speech synthesis: A review, *Proceedings of Eurospeech* (Vol. 1). Aalborg, Denmark.

Seppanen, T., Vayrynen, E., & Toivanen, J. (2003) Prosody-based classification of emotions in spoken Finnish, *8th European Conference on Speech Communication and Technology*. Geneval, Switzerland.

Sharma, R., Neumann, U., & Kim, C. (2002) Emotion recognition in spontaneous emotional utterances from movie sequences, *WSEAS International Conference on Electronics, Control, and Signal Processing*.

Shimura, Y., & Imaizumi, S. (1994) Infant's expression and perception of emotion through vocalizations, *Third International Conference on Spoken Language Processing*. Yokohama, Japan.

Simpson, J. A., Collins, W. A., Tran, S., & Haydon, K. C. (2007) Attachment and the Experience and Expression of Emotions in Romantic Relationships: A Developmental Perspective. *Journal of Personality and Social Psychology 92*(2), 355-367.

Slaney, M., & McRoberts, G. (1998) Baby Ears: A Recognition System for Affective Vocalizations, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Seattle, WA.

Smith, J. O. (2007) *Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications*: W3K Publishing.

Soltis, J. (2009) Vocal production, affect expression, and communicative function in African elephant rumble vocalizations. *Journal of the Acoustical Society of America 126*(4), 2210-2210.

Soltis, J., Leong, K., & Savage, A. (2005) African elephant vocal communication II: rumble variation reflects the individual identity and emotional state of callers. *Animal Behaviour 70*(3), 589-599.

Sroufe, L. A., Schork, E., Motti, E., Lawroski, N., & LaFreniere, P. (1984) The role of affect in emerging social competence. *Emotion, cognition, and behavior*, 289-319.

Steiniger, S., Schiel, F., Dioubina, O., & Raubold, S. (2002) Development of user-state conventions for the multimodal corpus in SmartKom, *Proc. Workshop on Multimodal Resources and Multimodal Systems* (pp. 33-37).

Stibbard, R. M. (2000) Automated extraction of ToBI annotation data from the Reading/Leeds emotional speech corpus, *Proceedings of the ISCA ITRW on Speech and Emotion* (pp. 60-65). Newcastle.

Strongman, K. T. (1996) The psychology of emotion. In *Theories of Emotion in Perspective*. New York, NY: John Wiley & Sons.

Sundberg, J. (1977) The acoustics of the singing voice. *Scientific American 231*(3), 82-91.

Symons, D. K. (2001) A dyad-oriented approach to distress and mother-child relationship outcomes in the first 24 months. *Parenting 1*(1-2), 101-122.

Tapus, A., Mataric, M., & Scassellati, B. (2007) The grand challenges in socially assistive robotics. *IEEE Robotics and Automation Magazine 14*(1), 35-42.

Tato, R., Santos, R., Kompe, R., & Pardo, J. M. (2002) Emotional space improves emotion recognition, *International Conference on Spoken Language Processing* (pp. 2029-2032). Denver, CO.

The MathWorks, I. Partial Least Squares Regression and Pricipal Components Regression Demo. http://www.mathworks.com/products/statistics/demos.html?file=/products/demos/shipping/stats/plspcrdemo.html

Tobias, R. D. (1995) An Introduction to Partial Least Squares Regression, *SUGI Proceedings*.

Trevarthen, C. (1990) Signs before speech. In T. A. Sebeok & J. U. Sebeok (Eds.), *The semiotic web*. Berlin: FRG: Mouton de Gruyter.

Trevarthen, C. (1993) The function of emotions in early infant communication and development. In J. Nadel & L. Camaioni (Eds.), *New Perspectives in Early Communicative Development* (pp. 48-81). London: Routledge.

Truong, K. P., Neerincx, M. A., & van Leeuwen, D. A. (2008) Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data, *Interspeech* (pp. 257-260). Brisbane, Australia.

Truong, K. P., & van Leeuwen, D. A. (2007) An 'open-set' detection evaluation methodology for automatic emotion recognition in speech, *Workshop on Paralinguistic Speech - between models and data* (pp. 5-10).

Ullakonoja, R. (2010) Pitch contours in Russian yes/no questions by Finns, *Speech Prosody* (Vol. 100072, pp. 1-4). Chicago, IL.

Varallyay Jr, G., Benyo, Z., & Illenyi, A. (2007) The development of the melody of the infant cry to detect disorders during infancy, *Proceedings of the Fifth IASTED International Conference on Biomedical Engineering*. Innsbruck, Austria.

Varallyay Jr, G., Benyo, Z., Illenyi, A., Farkas, Z., & Kovacs, L. (2004) Acoustic analysis of the infant cry: classical and new methods, *Proc. 26th Conf. IEEE Engineering in Medicine and Biology* (pp. 313-316). San Francisco, CA.

Ververidis, D., & Kotropoulos, C. (2006) Emotional speech recognition: Resources, features, and methods. *Speech Communication 48*(9), 1162-1181.

Vidrascu, L., & Devillers, L. (2005) Detection of real-life emotions in call centers, *Proc. Interspeech* (pp. 1841-1844).

Vihman, M. M., Macken, M. A., Miller, R., Simmons, H., & Miller, J. (1985) From babbling to speech: A re-assessment of the continuity issue. *Language 61*, 397-445.

Volkmar, F. R., Lord, C., Bailey, A., Schultz, R. T., & Klin, A. (2004) Autism and pervasive developmental disorders. *Journal of Child Psychology and Psychiatry 45*(1), 135-170.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005) Development of vocal tract length during early childhood: A magnetic

resonance imaging study. *Journal of the Acoustical Society of America 117*(1), 338-350.

Vosoughi, S., Roy, B., Frank, M. C., & Roy, D. (2010) Effects of Caretaker Prosody on Child Language Acquisition, *Speech Prosody* (Vol. 100120). Chicago, IL.

Warlaumont, A. S., Oller, D. K., & Buder, E. H. (2010) Data-driven automated acoustic analysis of human infant vocalizations using neural network tools. *Journal of the Acoustical Society of America 127*(4), 2563-2577.

Wasz-Hockert, O., Michelsson, K., & Lind, J. (1985) Twenty-five years of scandinavian cry research. In *Infant Crying: Theoretical and Research Perspectives*. New York, NY: Plenum Press.

Wermke, K., Mende, W., Manfredi, C., & Bruscaglioni, P. (2002) Developmental aspects of infant's cry melody and formants. *Med Eng Phys 24*(7-8), 501-514.

Wold, H. (1982) Soft Modeling: The Basic Design and Some Extensions. In H. Wold & K. G. Joreskog (Eds.), *Systems Under Indirect Observations: Causality, Structure, Prediction*. Amsterdam: Elsevier.

Wolfe, C. D., & Bell, M. A. (2007) The integration of cognition and emotion during infancy and early childhood: Regulatory processes associated with the development of working memory. *Brain and Cognition*.

Yeniay, O., & Goktas, A. (2002) A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics 31*, 99-111.

Yumoto, E., Sasaki, Y., & Okamura, H. (1984) Harmonics-to-Noise Ratio and Psychophysical Measurement of the Degree of Hoarseness. *Journal of Speech and Hearing Research 27*, 2-6.

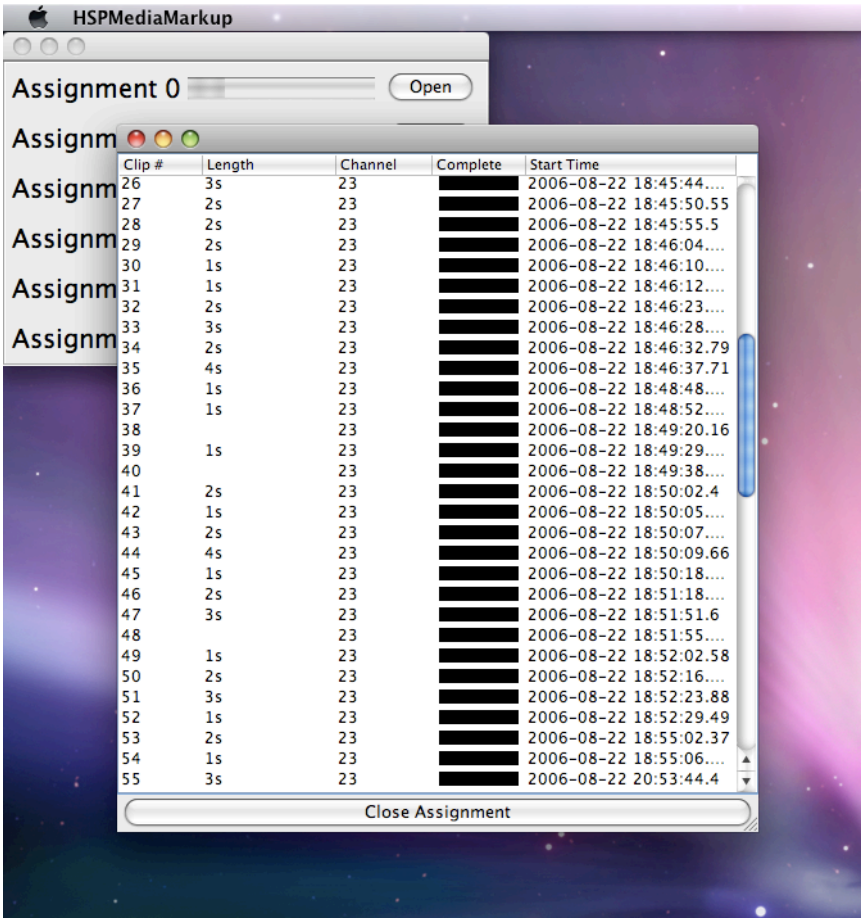
Zeifman, D. M. (2005) Crying behaviour and its impact on psychosocial child development: comment on Stifter, and Zeskind. In *Encyclopedia on Early Childhood Development [online]*. Montreal, Quebec: Centre of Excellence for Early Childhood Development.

Zeskind, P. S. (2005) Impact of the Cry of the Infant at Risk on Psychosocial Development. In R. E. Tremblay, R. G. Barr & R. D. Peters (Eds.), *Encyclopedia on Early Childhood Development* (pp. 1-7). Montreal, Quebec: Center of Excellence for Early Childhood Development.

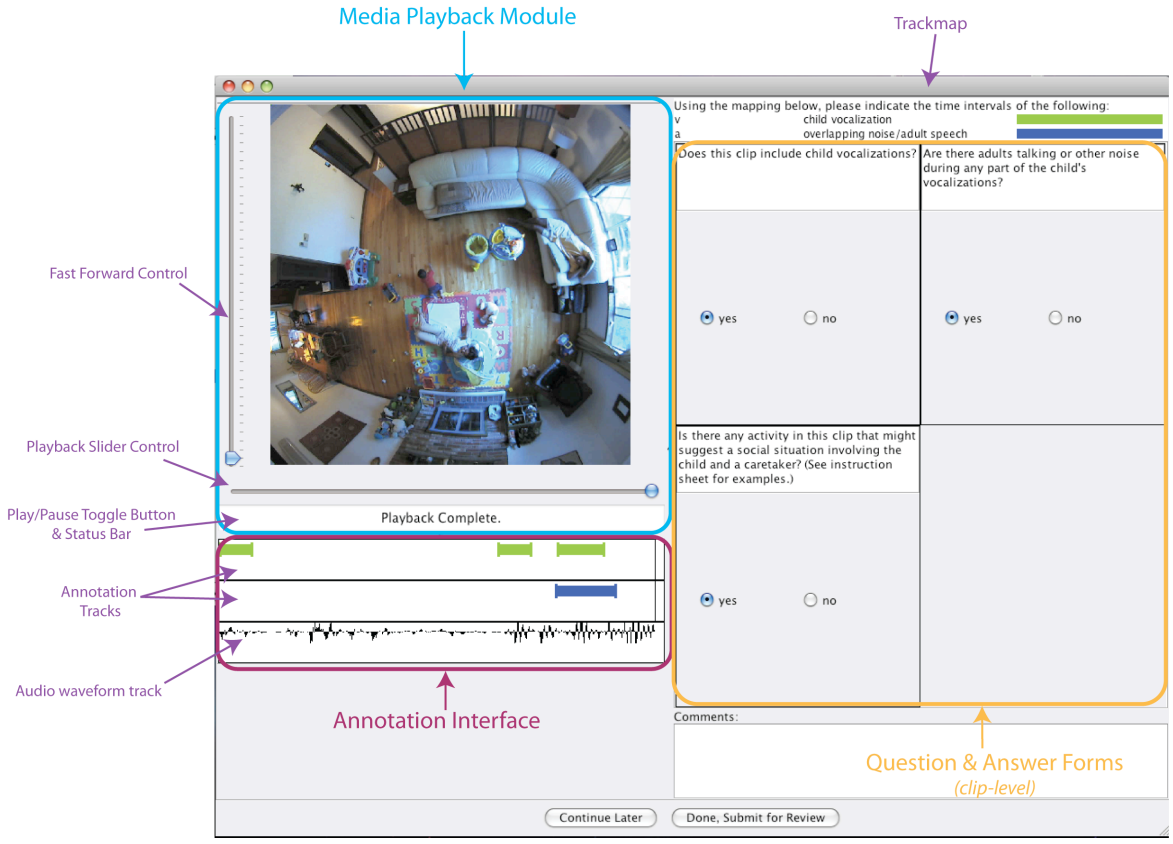
Zwaigenbaum, L., Bryson, S., Lord, C., Rogers, S., Carter, A., Carver, L., et al. (2009) Clinical assessment and management of toddlers with suspected Autism Spectrum Disorder: Insights from studies of high-risk infants. *Pediatrics 123*(5), 1383-1391.

Appendix A

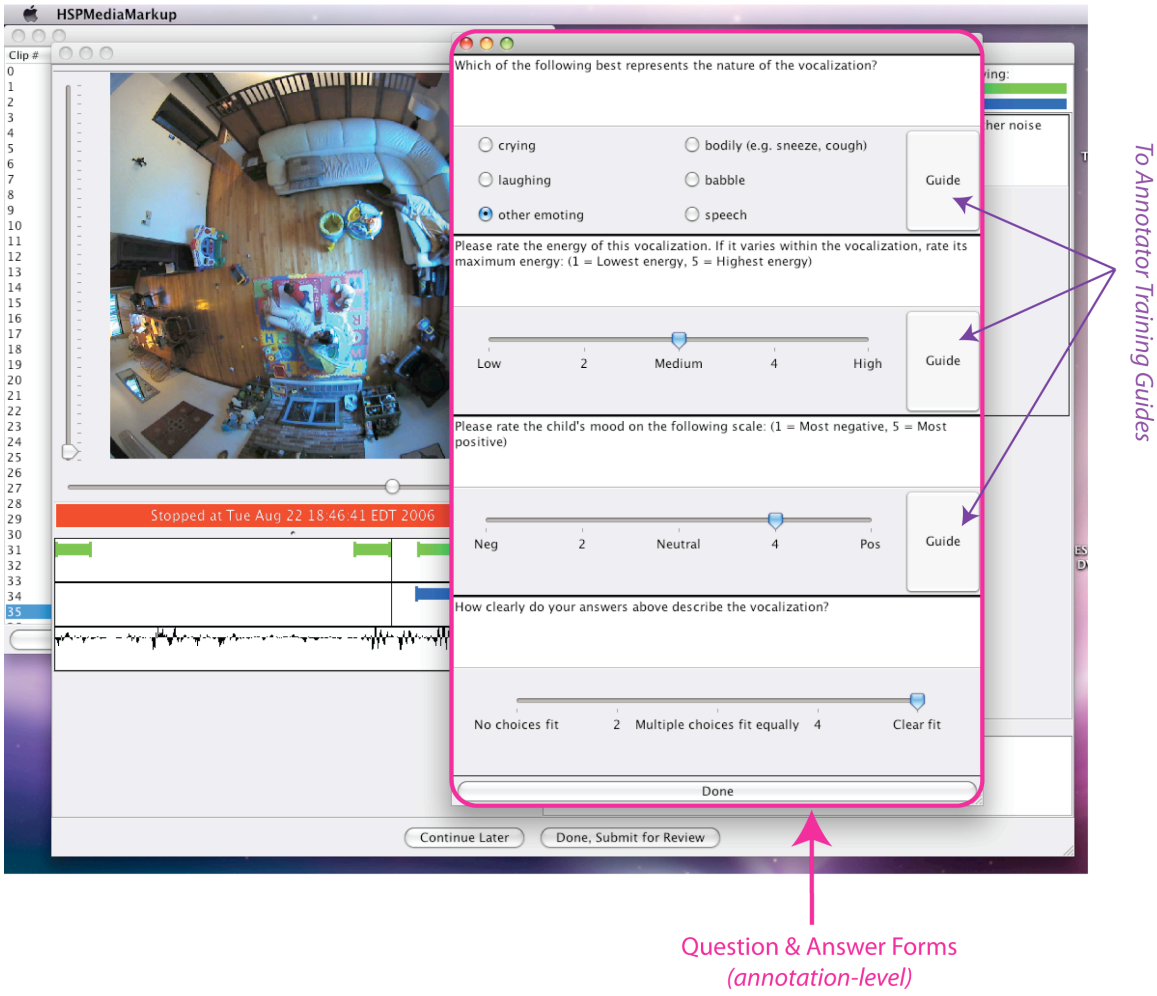
Interface Screenshots



Appendix A, Figure 1. Playback Browser List.



Appendix A, Figure 2. Playback and Annotation Interface (PAI)



Appendix A, Figure 3. Annotation-level Question & Answer Popup Dialog

Appendix A, Figure 4. Browsing the Annotator Training Guide

Appendix B

Question Configuration File

```
Multichoice;Does this clip include child vocalizations?;yes^no
Multichoice;Are there adults talking or other noise during any part of the child's
    vocalizations?;yes^no
Multichoice;Is there any activity in this clip that might suggest a social situation involving
    the child and a caretaker? (See instruction sheet for examples.);yes^no
Multichoice_Ann_v;Which of the following best represents the nature of the
    vocalization?;crying^bodily (e.g. sneeze, cough)^laughing^babble^other
    emoting^speech
Scale_Ann_v;Please rate the energy of this vocalization. If it varies within the
    vocalization, rate its maximum energy: (1 = Lowest energy, 5 = Highest
    energy);Low@1 - 2@2 - Medium@3 - 4@4 - High@5;3
Scale_Ann_v;Please rate the child's mood on the following scale: (1 = Most negative, 5 =
    Most positive);Neg@1 - 2@2 - Neutral@3 - 4@4 - Pos@5;3
Scale_Ann_v;How clearly do your answers above describe the vocalization?;No choices
    fit@1 - 2@2 - Multiple choices fit equally@3 - 4@4 - Clear fit@5;5
Trackmap;Using the mapping below, please indicate the time intervals of the
    following:;v,Child Vocalization,Green,true ~ a,Overlapping Noise/Adult
    Speech,Blue,true
```

Appendix C

Annotator Instruction Sheet

Definitions

Crying - in a child, it is an inarticulate, often prolonged expression of a negative state, and can range from soft weeping to screaming, depending on the energy of expression. It can also begin with whining or other fussy vocalizations.

Laughing - expressing certain positive emotions, especially amusement or delight, by a series of spontaneous, usually unarticulated sounds, such as heehee, hehe, haha.

Bodily - a vocalization that is produced by reflexive bodily functions and therefore carries no emotional content, such as a sneeze, cough, or burp.

Babble - a vocalization where the child is attempting to express speech, with or without emotion. It must include at least two clearly articulated consonant-vowels, i.e. at least two clear syllables.

Speech - the child is clearly speaking in articulated, recognizable words.

Other Emoting - any other vocalization uttered by the child.

Rules of Thumb

If you are undecided between...

... crying vs. other emoting -- choose **OTHER EMOTING**

... laughing vs. other emoting -- choose **OTHER EMOTING**

... babble vs. emoting -- listen for the number of clear consonant-vowels (i.e. syllables) in the vocalization. If there are two or more, then it's babble. **Less than two syllables is emoting.**

Social Situation Examples

Here are some example events to look for to determine a possible social situation/interaction between the child and caretaker. **If at least one of these events occur in a clip, answer "yes" to the social situation question.**

Dialog between child and caretaker
Caretaker talking to child
Child approaching caretaker
Caretaker approaching child
Caretaker pointing
Child pointing
Close proximity between child and caretaker

Appendix D

Per-Annotator Agreement

Table D - 1. Per-Annotator Agreement for Social Question

Annotator	# vocalizations annotated	# vocalizations agreed	% agreed
1	4834	4166	86.2
2	3314	2829	85.4
3	1063	401	**37.7
4	3226	2846	88.2
5	3247	2854	87.9
6	2342	1868	79.8
7	3030	2360	77.9

** The performance of Annotator #3 is clearly an outlier in the Social question. Given this anomaly, all assignments completed by this annotator were reassigned among the other annotators. For each of these assignments, care was taken to make sure that the annotator who initially shared an assignment with Annotator #3 did not receive it a second time.

Table D - 2. Per-Annotator Agreement for Nature Question

Annotator	# vocalizations annotated	# vocalizations agreed	% agreed
1	4807	3356	69.8
2	3304	2318	70.2
3	1071	728	68.0
4	3202	2281	71.2
5	3234	2276	70.4
6	2331	1679	72.0
7	3051	2186	71.6

Table D - 3. Per-Annotator Agreement for Energy Question

Annotator	# vocalizations annotated	# vocalizations agreed	% agreed	# disagreements by 1 pt	% agreed by ≤1 pt
1	4834	2451	50.7	2045	93.0
2	3325	1700	51.1	1406	93.4

3	1076	433	40.2	506	87.3
4	3226	1729	53.6	1320	94.5
5	3250	1645	50.6	1366	92.6
6	2343	949	40.5	1051	85.4
7	3112	1322	42.5	1379	86.8

Table D - 4. Per-Annotator Agreement for Mood Question

Annotator	# vocalizations annotated	# vocalizations agreed	% agreed	# disagreements by 1 pt	% agreed by ≤1 pt
1	4833	2283	47.2	2376	96.4
2	3325	1652	49.7	1472	94.0
3	1076	492	45.7	508	92.9
4	3227	1549	48.0	1494	94.3
5	3252	1686	51.8	1383	94.4
6	2342	1329	56.7	913	95.7
7	3075	1634	53.1	1327	96.3

Appendix E

Praat Script Examples

E.1. Praat Script for a Child Vocalization

This example is for a pruned child vocalization starting at 1183770446787 milliseconds, Unix epoch time (July 6, 2007 21:07:26.787 EDT) and ending at 1183770448349 milliseconds, Unix epoch time (July 6, 2007 21:07:28.349 EDT).

```
Read from file... /Users/sophia/prosody/wavs/childwavs/prunedvoc/pv_1183770446787_1183770448349.wav
printline starting file  9200 of 9202: pv_1183770446787_1183770448349
select Sound pv_1183770446787_1183770448349
dur = Get total duration
printline duration = 'dur:6'
select Sound pv_1183770446787_1183770448349
To Intensity... 100 0 yes
select Intensity pv_1183770446787_1183770448349
imean = Get mean... 0 0 energy
imax = Get maximum... 0 0 Parabolic
imin = Get minimum... 0 0 Parabolic
istdev = Get standard deviation... 0 0
Write to text file...
/Users/sophia/prosody/wavs/childwavs/features/pv_1183770446787_1183770448349.Intensity

select Sound pv_1183770446787_1183770448349
To Pitch... 0.01 75 2000
select Pitch pv_1183770446787_1183770448349
Kill octave jumps
Interpolate
Smooth... 10
f0mean = Get mean... 0 0 Hertz
f0max = Get maximum... 0 0 Hertz Parabolic
f0min = Get minimum... 0 0 Hertz Parabolic
f0slope2 = Get mean absolute slope... Semitones
f0stdev = Get standard deviation... 0 0 Hertz

Write to text file...
/Users/sophia/prosody/wavs/childwavs/features/pv_1183770446787_1183770448349.Pitch

select Sound pv_1183770446787_1183770448349
manip = To Manipulation... 0.01 75 2000
Extract pitch tier
Rename... pt_pv_1183770446787_1183770448349
Stylize... 2.0 Semitones
plus manip
Replace pitch tier
select PitchTier pt_pv_1183770446787_1183770448349
```

```
numpts = Get number of points
Write to headerless spreadsheet file...
/Users/sophia/prosody/wavs/childwavs/features/pv_1183770446787_1183770448349.Stylized
select Sound pv_1183770446787_1183770448349
To Spectrum... yes
fft_centroid1 = Get centre of gravity... 1
fft_centroid2 = Get centre of gravity... 2
fftstd1 = Get standard deviation... 1
fftstd2 = Get standard deviation... 2
fftskew1 = Get skewness... 1
fftskew2 = Get skewness... 2
fftkurt1 = Get kurtosis... 1
fftkurt2 = Get kurtosis... 2

To Ltas... 125
ltas_freqofmax = Get frequency of maximum... 0 0 none
ltas_freqofmin = Get frequency of minimum... 0 0 None
ltas_max = Get maximum... 0 0 None
ltas_mean = Get mean... 0 0 energy
ltas_stdev = Get standard deviation... 0 0 energy
ltas_min = Get minimum... 0 0 None

select Sound pv_1183770446787_1183770448349
To Harmonicity (cc)... 0.03 75 0.1 4.5
hnr_min = Get minimum... 0 0 Parabolic
hnr_mean = Get mean... 0 0
hnr_max = Get maximum... 0 0 Parabolic
hnr_std = Get standard deviation... 0 0
hnr_timeofmax = Get time of maximum... 0 0 Parabolic
hnr_timeofmin = Get time of minimum... 0 0 Parabolic

select Sound pv_1183770446787_1183770448349
To MFCC... 16 0.015 0.005 100.0 100.0 0.0
To Matrix
Transpose
To TableOfReal
mfcc1mean = Get column mean (index)... 1
mfcc1std = Get column stdev (index)... 1
mfcc2mean = Get column mean (index)... 2
mfcc2std = Get column stdev (index)... 2
mfcc3mean = Get column mean (index)... 3
mfcc3std = Get column stdev (index)... 3
mfcc4mean = Get column mean (index)... 4
mfcc4std = Get column stdev (index)... 4
mfcc5mean = Get column mean (index)... 5
mfcc5std = Get column stdev (index)... 5
mfcc6mean = Get column mean (index)... 6
mfcc6std = Get column stdev (index)... 6
mfcc7mean = Get column mean (index)... 7
mfcc7std = Get column stdev (index)... 7
mfcc8mean = Get column mean (index)... 8
```

```
mfcc8std = Get column stdev (index)... 8
mfcc9mean = Get column mean (index)... 9
mfcc9std = Get column stdev (index)... 9
mfcc10mean = Get column mean (index)... 10
mfcc10std = Get column stdev (index)... 10
mfcc11mean = Get column mean (index)... 11
mfcc11std = Get column stdev (index)... 11
mfcc12mean = Get column mean (index)... 12
mfcc12std = Get column stdev (index)... 12
mfcc13mean = Get column mean (index)... 13
mfcc13std = Get column stdev (index)... 13
mfcc14mean = Get column mean (index)... 14
mfcc14std = Get column stdev (index)... 14
mfcc15mean = Get column mean (index)... 15
mfcc15std = Get column stdev (index)... 15
mfcc16mean = Get column mean (index)... 16
mfcc16std = Get column stdev (index)... 16

select Sound pv_1183770446787_1183770448349
To Formant (burg)... 0.0 5 8000 0.025 50
fmt1min = Get minimum... 1 0 0 Hertz Parabolic
fmt1mean = Get mean... 1 0 0 Hertz
fmt1max = Get maximum... 1 0 0 Hertz Parabolic
fmt1std = Get standard deviation... 1 0 0 Hertz
fmt2min = Get minimum... 2 0 0 Hertz Parabolic
fmt2mean = Get mean... 2 0 0 Hertz
fmt2max = Get maximum... 2 0 0 Hertz Parabolic
fmt2std = Get standard deviation... 2 0 0 Hertz
fmt3min = Get minimum... 3 0 0 Hertz Parabolic
fmt3mean = Get mean... 3 0 0 Hertz
fmt3max = Get maximum... 3 0 0 Hertz Parabolic
fmt3std = Get standard deviation... 3 0 0 Hertz
fmt4min = Get minimum... 4 0 0 Hertz Parabolic
fmt4mean = Get mean... 4 0 0 Hertz
fmt4max = Get maximum... 4 0 0 Hertz Parabolic
fmt4std = Get standard deviation... 4 0 0 Hertz
fmt5min = Get minimum... 5 0 0 Hertz Parabolic
fmt5mean = Get mean... 5 0 0 Hertz
fmt5max = Get maximum... 5 0 0 Hertz Parabolic
fmt5std = Get standard deviation... 5 0 0 Hertz
fileappend "/Users/sophia/prosody/wavs/childwavs/features/allfeatures_extended.txt"
pv_1183770446787_1183770448349.wav
'tab$''imin:6''tab$''imean:6''tab$''imax:6''tab$''istdev:6''tab$''f0slope2:6''tab$''f0stdev:6''tab$''f0min
:6''tab$''f0mean:6''tab$''f0max:6''tab$''fft_centroid1:6''tab$''fft_centroid2:6''tab$''fftstd1:6''tab$''
fftstd2:6''tab$''fftskew1:6''tab$''fftskew2:6''tab$''fftkurt1:6''tab$''fftkurt2:6''tab$''ltas_freqofmax:
6''tab$''ltas_freqofmin:6''tab$''ltas_max:6''tab$''ltas_mean:6''tab$''ltas_stdev:6''tab$''ltas_min:6''tab
$''hnr_min:6''tab$''hnr_mean:6''tab$''hnr_max:6''tab$''hnr_std:6''tab$''hnr_timeofmax:6''tab$''hnr_ti
meofmin:6''tab$''mfcc1mean:6''tab$''mfcc1std:6''tab$''mfcc2mean:6''tab$''mfcc2std:6''tab$''mfcc3mea
n:6''tab$''mfcc3std:6''tab$''mfcc4mean:6''tab$''mfcc4std:6''tab$''mfcc5mean:6''tab$''mfcc5std:6''tab
$''mfcc6mean:6''tab$''mfcc6std:6''tab$''mfcc7mean:6''tab$''mfcc7std:6''tab$''mfcc8mean:6''tab$''mfc
c8std:6''tab$''mfcc9mean:6''tab$''mfcc9std:6''tab$''mfcc10mean:6''tab$''mfcc10std:6''tab$''mfcc11mea
```

n:6''tab$''mfcc11std:6''tab$''mfcc12mean:6''tab$''mfcc12std:6''tab$''mfcc13mean:6''tab$''mfcc13std:6''tab$''mfcc14mean:6''tab$''mfcc14std:6''tab$''mfcc15mean:6''tab$''mfcc15std:6''tab$''mfcc16mean:6''tab$''mfcc16std:6''tab$''fmt1min:6''tab$''fmt1mean:6''tab$''fmt1max:6''tab$''fmt1std:6''tab$''fmt2min:6''tab$''fmt2mean:6''tab$''fmt2max:6''tab$''fmt2std:6''tab$''fmt3min:6''tab$''fmt3mean:6''tab$''fmt3max:6''tab$''fmt3std:6''tab$''fmt4min:6''tab$''fmt4mean:6''tab$''fmt4max:6''tab$''fmt4std:6''tab$''fmt5min:6''tab$''fmt5mean:6''tab$''fmt5max:6''tab$''fmt5std:6''tab$''numpts''newline$'

E.2. Example Praat Script for Adult Speech

This particular Praat script is for all adult speech occurring within a 30 second window after the pruned child vocalization starting at 1147790847054 milliseconds, Unix epoch time (May 16, 2006 10:47:27.054 EDT) and ending at 1147790847440 milliseconds, Unix epoch time (May 16, 2006 10:47:27.440 EDT). Three fragments of adult speech are combined to form this 30-second window – 74ms, 46ms, and 1410ms in duration – with three periods of Silence synthesized to fill in the gaps –18.58 sec, 1.83 sec, and 8.07 sec in length.

```
Create Sound from formula... silence_voc_1147790847440_1147790866020 Mono 0.0 18.58 48000 0
printline reading from input wav
Read from file...
/Users/sophia/prosody/wavs/adultwavs/pafter30/adaft_1147790866020_1147790866094.wav
Create Sound from formula... silence_voc_1147790866094_1147790867925 Mono 0.0 1.831 48000 0
printline reading from input wav
Read from file...
/Users/sophia/prosody/wavs/adultwavs/pafter30/adaft_1147790867925_1147790867971.wav
Create Sound from formula... silence_voc_1147790867971_1147790876040 Mono 0.0 8.069 48000 0
printline reading from input wav
Read from file...
/Users/sophia/prosody/wavs/adultwavs/pafter30/adaft_1147790876040_1147790877450.wav


printline concatenating input wavs and silence
select Sound silence_voc_1147790847440_1147790866020
plus Sound adaft_1147790866020_1147790866094
printline concatenating input wavs and silence
plus Sound silence_voc_1147790866094_1147790867925
plus Sound adaft_1147790867925_1147790867971
printline concatenating input wavs and silence
plus Sound silence_voc_1147790867971_1147790876040
plus Sound adaft_1147790876040_1147790877450
Concatenate
Copy... adaft4voc_1147790847054_1147790847440
Write to WAV file...
/Users/sophia/prosody/wavs/adultwavs/pafter30/may06/outwavs/adaft4voc_1147790847054_1147790847440.wav
printline finished concatenating


printline adaft4voc_1147790847054_1147790847440
printline computing intensity
select Sound adaft4voc_1147790847054_1147790847440
dur = Get total duration
printline duration = 'dur:6'
select Sound adaft4voc_1147790847054_1147790847440
To Intensity... 100 0 yes
select Intensity adaft4voc_1147790847054_1147790847440
imean = Get mean... 0 0 energy
```

```
imax = Get maximum... 0 0 Parabolic
imin = Get minimum... 0 0 Parabolic
istdev = Get standard deviation... 0 0
Write to text file...
/Users/sophia/prosody/wavs/adultwavs/pafter30/may06/features/adaft4voc_1147790847054_1147790847440.Intensity

printline computing pitch
select Sound adaft4voc_1147790847054_1147790847440
To Pitch... 0.01 75 2000
select Pitch adaft4voc_1147790847054_1147790847440
Kill octave jumps
Interpolate
Smooth... 10
f0mean = Get mean... 0 0 Hertz
f0max = Get maximum... 0 0 Hertz Parabolic
f0min = Get minimum... 0 0 Hertz Parabolic
f0slope2 = Get mean absolute slope... Semitones
f0stdev = Get standard deviation... 0 0 Hertz

Write to text file...
/Users/sophia/prosody/wavs/adultwavs/pafter30/may06/features/adaft4voc_1147790847054_1147790847440.Pitch

printline computing stylization
select Sound adaft4voc_1147790847054_1147790847440
manip = To Manipulation... 0.01 75 2000
Extract pitch tier
Rename... pt_adaft4voc_1147790847054_1147790847440
Stylize... 2.0 Semitones
plus manip
Replace pitch tier
select PitchTier pt_adaft4voc_1147790847054_1147790847440
numpts = Get number of points
Write to headerless spreadsheet file...
/Users/sophia/prosody/wavs/adultwavs/pafter30/may06/features/adaft4voc_1147790847054_1147790847440.Stylized

printline computing fft
select Sound adaft4voc_1147790847054_1147790847440
To Spectrum... yes
fft_centroid1 = Get centre of gravity... 1
fft_centroid2 = Get centre of gravity... 2
fftstd1 = Get standard deviation... 1
fftstd2 = Get standard deviation... 2
fftskew1 = Get skewness... 1
fftskew2 = Get skewness... 2
fftkurt1 = Get kurtosis... 1
fftkurt2 = Get kurtosis... 2

printline computing Ltas
```

```
To Ltas... 125
ltas_freqofmax = Get frequency of maximum... 0 0 none
ltas_freqofmin = Get frequency of minimum... 0 0 None
ltas_max = Get maximum... 0 0 None
ltas_mean = Get mean... 0 0 energy
ltas_stdev = Get standard deviation... 0 0 energy
ltas_min = Get minimum... 0 0 None

printline computing hnr
select Sound adaft4voc_1147790847054_1147790847440
To Harmonicity (cc)... 0.03 75 0.1 4.5
hnr_min = Get minimum... 0 0 Parabolic
hnr_mean = Get mean... 0 0
hnr_max = Get maximum... 0 0 Parabolic
hnr_std = Get standard deviation... 0 0
hnr_timeofmax = Get time of maximum... 0 0 Parabolic
hnr_timeofmin = Get time of minimum... 0 0 Parabolic

printline computing mfcc's
select Sound adaft4voc_1147790847054_1147790847440
To MFCC... 16 0.015 0.005 100.0 100.0 0.0
To Matrix
Transpose
To TableOfReal
mfcc1mean = Get column mean (index)... 1
mfcc1std = Get column stdev (index)... 1
mfcc2mean = Get column mean (index)... 2
mfcc2std = Get column stdev (index)... 2
mfcc3mean = Get column mean (index)... 3
mfcc3std = Get column stdev (index)... 3
mfcc4mean = Get column mean (index)... 4
mfcc4std = Get column stdev (index)... 4
mfcc5mean = Get column mean (index)... 5
mfcc5std = Get column stdev (index)... 5
mfcc6mean = Get column mean (index)... 6
mfcc6std = Get column stdev (index)... 6
mfcc7mean = Get column mean (index)... 7
mfcc7std = Get column stdev (index)... 7
mfcc8mean = Get column mean (index)... 8
mfcc8std = Get column stdev (index)... 8
mfcc9mean = Get column mean (index)... 9
mfcc9std = Get column stdev (index)... 9
mfcc10mean = Get column mean (index)... 10
mfcc10std = Get column stdev (index)... 10
mfcc11mean = Get column mean (index)... 11
mfcc11std = Get column stdev (index)... 11
mfcc12mean = Get column mean (index)... 12
mfcc12std = Get column stdev (index)... 12
mfcc13mean = Get column mean (index)... 13
mfcc13std = Get column stdev (index)... 13
mfcc14mean = Get column mean (index)... 14
```

```
mfcc14std = Get column stdev (index)... 14
mfcc15mean = Get column mean (index)... 15
mfcc15std = Get column stdev (index)... 15
mfcc16mean = Get column mean (index)... 16
mfcc16std = Get column stdev (index)... 16

printline computing formants
select Sound adaft4voc_1147790847054_1147790847440
To Formant (burg)... 0.0 5 8000 0.025 50
fmt1min = Get minimum... 1 0 0 Hertz Parabolic
fmt1mean = Get mean... 1 0 0 Hertz
fmt1max = Get maximum... 1 0 0 Hertz Parabolic
fmt1std = Get standard deviation... 1 0 0 Hertz
fmt2min = Get minimum... 2 0 0 Hertz Parabolic
fmt2mean = Get mean... 2 0 0 Hertz
fmt2max = Get maximum... 2 0 0 Hertz Parabolic
fmt2std = Get standard deviation... 2 0 0 Hertz
fmt3min = Get minimum... 3 0 0 Hertz Parabolic
fmt3mean = Get mean... 3 0 0 Hertz
fmt3max = Get maximum... 3 0 0 Hertz Parabolic
fmt3std = Get standard deviation... 3 0 0 Hertz
fmt4min = Get minimum... 4 0 0 Hertz Parabolic
fmt4mean = Get mean... 4 0 0 Hertz
fmt4max = Get maximum... 4 0 0 Hertz Parabolic
fmt4std = Get standard deviation... 4 0 0 Hertz
fmt5min = Get minimum... 5 0 0 Hertz Parabolic
fmt5mean = Get mean... 5 0 0 Hertz
fmt5max = Get maximum... 5 0 0 Hertz Parabolic
fmt5std = Get standard deviation... 5 0 0 Hertz

printline appending features to file
fileappend
"/Users/sophia/prosody/wavs/adultwavs/pafter30/may06/features/allfeatures_extended_adaft_pruned_ma
y06.txt" adaft4voc_1147790847054_1147790847440
'tab$''imin:6''tab$''imean:6''tab$''imax:6''tab$''istdev:6''tab$''f0slope2:6''tab$''f0stdev:6''tab$''f0min
:6''tab$''f0mean:6''tab$''f0max:6''tab$''fft_centroid1:6''tab$''fft_centroid2:6''tab$''fftstd1:6''tab$''
fftstd2:6''tab$''fftskew1:6''tab$''fftskew2:6''tab$''fftkurt1:6''tab$''fftkurt2:6''tab$''ltas_freqofmax:
6''tab$''ltas_freqofmin:6''tab$''ltas_max:6''tab$''ltas_mean:6''tab$''ltas_stdev:6''tab$''ltas_min:6''tab
$''hnr_min:6''tab$''hnr_mean:6''tab$''hnr_max:6''tab$''hnr_std:6''tab$''hnr_timeofmax:6''tab$''hnr_ti
meofmin:6''tab$''mfcc1mean:6''tab$''mfcc1std:6''tab$''mfcc2mean:6''tab$''mfcc2std:6''tab$''mfcc3mea
n:6''tab$''mfcc3std:6''tab$''mfcc4mean:6''tab$''mfcc4std:6''tab$''mfcc5mean:6''tab$''mfcc5std:6''tab
$''mfcc6mean:6''tab$''mfcc6std:6''tab$''mfcc7mean:6''tab$''mfcc7std:6''tab$''mfcc8mean:6''tab$''mfc
c8std:6''tab$''mfcc9mean:6''tab$''mfcc9std:6''tab$''mfcc10mean:6''tab$''mfcc10std:6''tab$''mfcc11mea
n:6''tab$''mfcc11std:6''tab$''mfcc12mean:6''tab$''mfcc12std:6''tab$''mfcc13mean:6''tab$''mfcc13std:6''
tab$''mfcc14mean:6''tab$''mfcc14std:6''tab$''mfcc15mean:6''tab$''mfcc15std:6''tab$''mfcc16mean:6''ta
b$''mfcc16std:6''tab$''fmt1min:6''tab$''fmt1mean:6''tab$''fmt1max:6''tab$''fmt1std:6''tab$''fmt2min:6
''tab$''fmt2mean:6''tab$''fmt2max:6''tab$''fmt2std:6''tab$''fmt3min:6''tab$''fmt3mean:6''tab$''fmt3
max:6''tab$''fmt3std:6''tab$''fmt4min:6''tab$''fmt4mean:6''tab$''fmt4max:6''tab$''fmt4std:6''tab$''f
mt5min:6''tab$''fmt5mean:6''tab$''fmt5max:6''tab$''fmt5std:6''tab$''numpts''newline$'
```

Appendix F

Optimal Number of PLS Components

Table F-1 lists the optimal number of PLS Components for each of our overall (9-24 month period) models. These numbers were derived using a heuristic technique in which we plot Mean-Squared Error (MSE) as a function of components and choose the minimum value. In many cases, this graph of residual variation contains a clear global minimum beyond which MSE begins to rise with additional components.

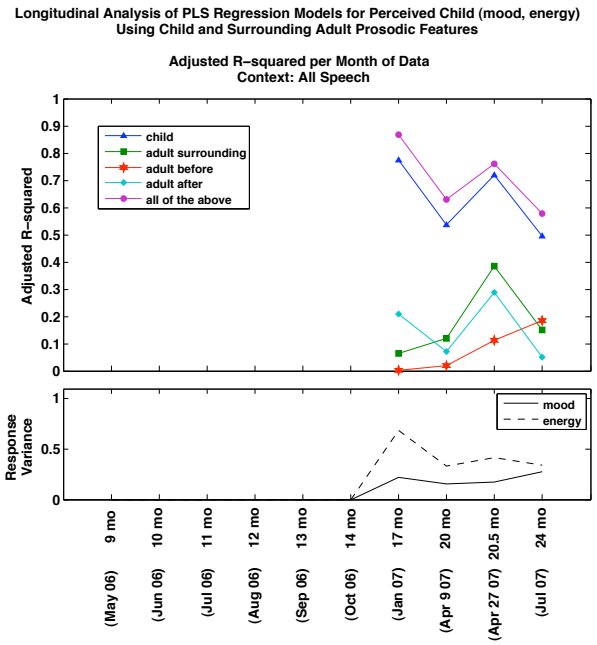
This listing is provided for illustration; we also apply the same heuristic to the monthly subsets for all the contexts but do not list them here.

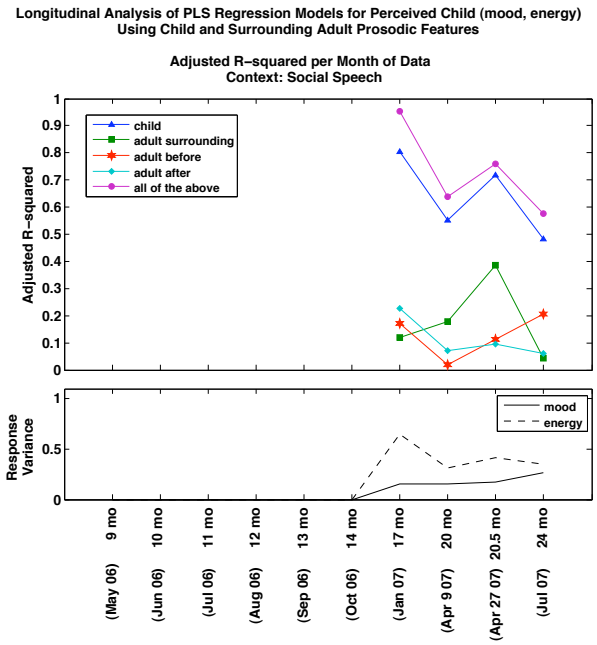
Table F- 1. Optimal Number of PLS Components for Overall Models

Context	Child	Adult			Combined
		Before	After	Surr	
All	24	19	50	50	25
Social Situations Only	16	15	17	14	16
Not Bodily	24	21	50	49	17
Social Not Bodily	15	15	17	16	14
Crying	5	4	3	3	5
Social Crying	3	4	5	4	4
Babble	4	4	4	1	6
Social Babble	4	4	4	2	6
Speech	5	3	3	3	5
Social Speech	6	2	1	3	6
Other Emoting	11	16	18	16	13
Social Other Emoting	10	17	15	11	12
Laughing	2	1	1	1	3
Social Laughing	2	1	1	2	3

Appendix G

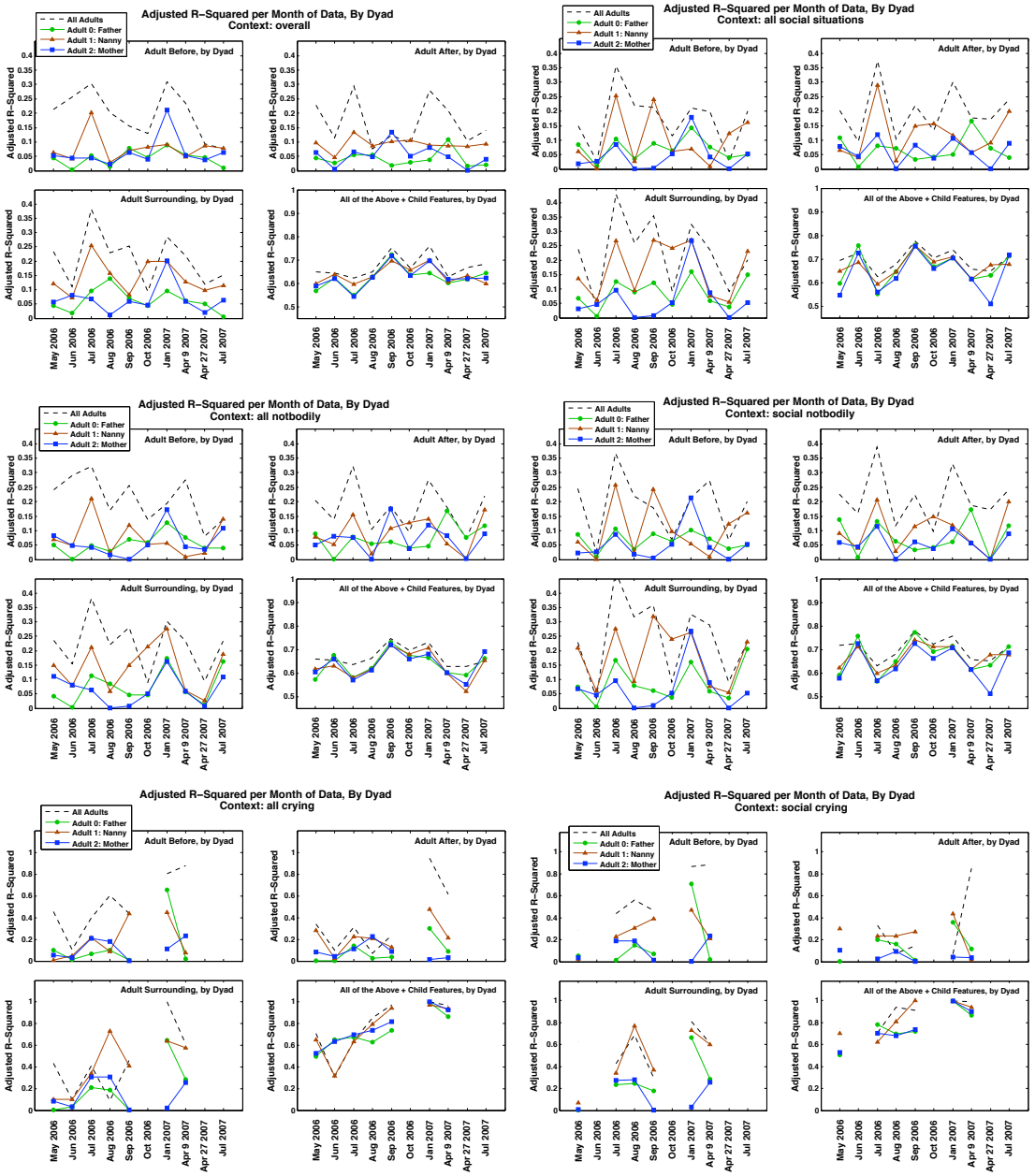
Supplementary Results

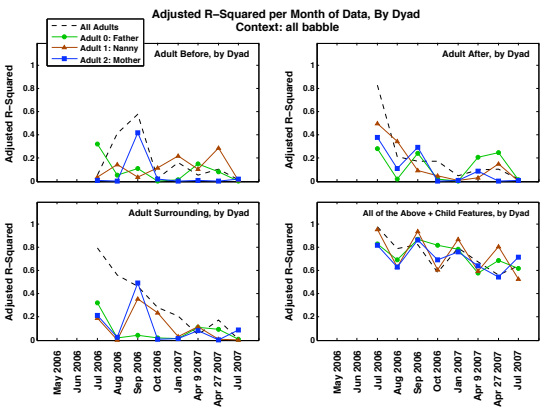


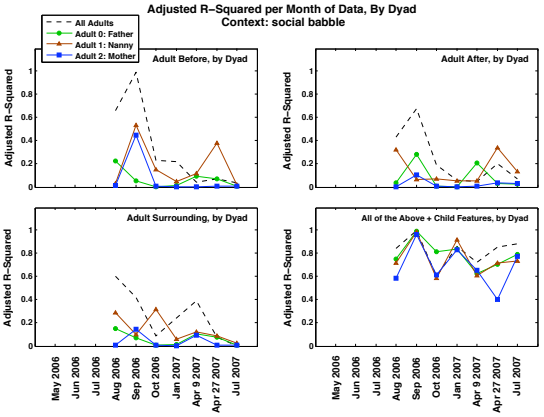


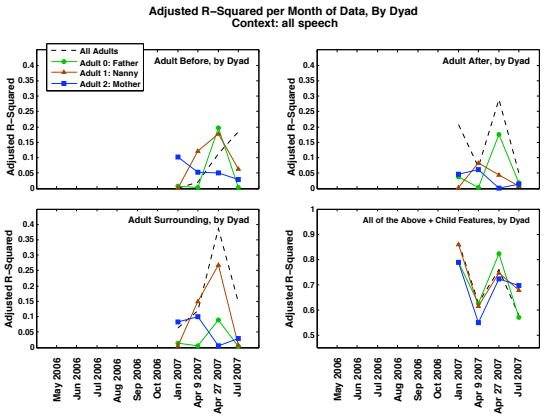
Appendix H

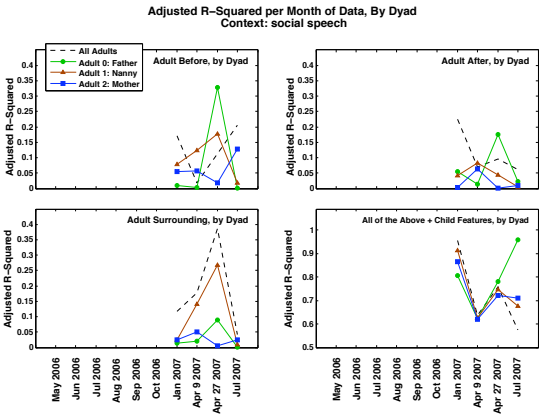
Longitudinal Analysis: Dyadic Results

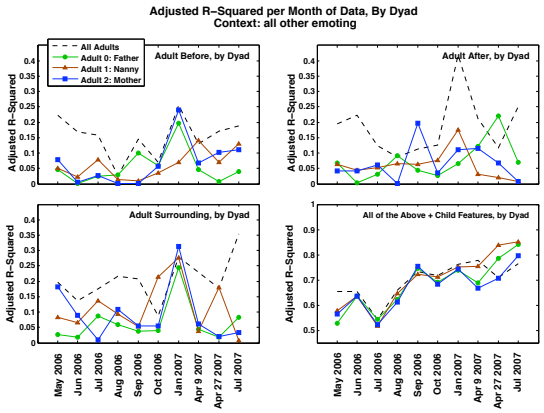


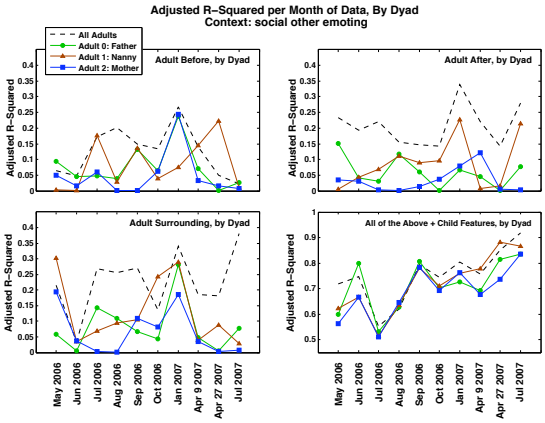

Adjusted R−Squared per Month of Data, By Dyad
Context: all babble
All Adults
Adult 0: Father
Adult 1: Nanny
Adult 2: Mother
Adult Before, by Dyad
Adult After, by Dyad
Adult Surrounding, by Dyad
All of the Above + Child Features, by Dyad
Adjusted R−Squared
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Jan 2007
Apr 9 2007
Apr 27 2007
Jul 2007


Adjusted R−Squared per Month of Data, By Dyad
Context: social babble
All Adults
Adult 0: Father
Adult 1: Nanny
Adult 2: Mother
Adult Before, by Dyad
Adult After, by Dyad
Adult Surrounding, by Dyad
All of the Above + Child Features, by Dyad
Adjusted R−Squared
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Jan 2007
Apr 9 2007
Apr 27 2007
Jul 2007


Adjusted R−Squared per Month of Data, By Dyad
Context: all speech
All Adults
Adult 0: Father
Adult 1: Nanny
Adult 2: Mother
Adult Before, by Dyad
Adult After, by Dyad
Adult Surrounding, by Dyad
All of the Above + Child Features, by Dyad
Adjusted R−Squared
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Jan 2007
Apr 9 2007
Apr 27 2007
Jul 2007


Adjusted R−Squared per Month of Data, By Dyad
Context: social speech
All Adults
Adult 0: Father
Adult 1: Nanny
Adult 2: Mother
Adult Before, by Dyad
Adult After, by Dyad
Adult Surrounding, by Dyad
All of the Above + Child Features, by Dyad
Adjusted R−Squared
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Jan 2007
Apr 9 2007
Apr 27 2007
Jul 2007


Adjusted R−Squared per Month of Data, By Dyad
Context: all other emoting
All Adults
Adult 0: Father
Adult 1: Nanny
Adult 2: Mother
Adult Before, by Dyad
Adult After, by Dyad
Adult Surrounding, by Dyad
All of the Above + Child Features, by Dyad
Adjusted R−Squared
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Jan 2007
Apr 9 2007
Apr 27 2007
Jul 2007


Adjusted R−Squared per Month of Data, By Dyad
Context: social other emoting
All Adults
Adult 0: Father
Adult 1: Nanny
Adult 2: Mother
Adult Before, by Dyad
Adult After, by Dyad
Adult Surrounding, by Dyad
All of the Above + Child Features, by Dyad
Adjusted R−Squared
May 2006
Jun 2006
Jul 2006
Aug 2006
Sep 2006
Oct 2006
Jan 2007
Apr 9 2007
Apr 27 2007
Jul 2007