USC-SIPI REPORT #413

Automatic Quantification and Prediction of Human Subjective Judgments In Behavioral Signal Processing

by

Matthew P. Black

February 2012

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

AUTOMATIC QUANTIFICATION AND PREDICTION OF HUMAN SUBJECTIVE JUDGMENTS IN BEHAVIORAL SIGNAL PROCESSING

by

Matthew P. Black

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

February 2012

Dedication

I dedicate this dissertation to my family, my fiancé Rachna, and her family. To my parents, I owe you everything. To my sisters, thank you for your friendship and for surrounding me with the perspective that every hopeless brother desperately needs.

I have had the good fortune and honor to know and learn from all four of my grandparents and two of my great-grandparents. To my Granny, thank you for providing me with the infrastructure for a solid technical background through repeated readings of *My First Picture Book of Numbers.* To my Grandpa, a brilliant optical physicist, you have been a wonderful role model. Your background and success helped give me the confidence to pursue and achieve one of my early career goals: a Ph.D. in science/engineering ... "It was nothing!"

To my future wife, Rachna, thank you for all your support during my tenure at the University of Southern California. You have been there every step of the way, celebrating the good times, and challenging me and encouraging me when the going got tough. You are the best friend that every person needs and the partner I always hoped I would find. I love you, and I cannot wait to marry you and become a member of your great and supportive family – I am truly blessed to have so many wonderful people in my life!

Acknowledgements

I would like to thank my advisor and mentor, Dr. Shrikanth S. Narayanan, with whom I have had the honor of studying under while at the University of Southern California. Your guidance has truly helped me, your leadership has inspired me, and I am very grateful for all you have done for me.

Many thanks also to my committee members, Dr. Antonio Ortega and Dr. Gayla Margolin, for your valuable insight and suggestions. I would also like to acknowledge the help from Dr. Panayiotis Georgiou and Dr. Sungbok Lee, key contributors in my work. Finally, I would like to thank all my fellow colleagues and labmates, whose collaborations, interactions, and friendships have been so critical in the development and completion of my thesis.

This dissertation was supported in part by grants from the following organizations: the National Science Foundation, the National Institute of Child Health and Human Development, the Viterbi Research Innovation Fund, Autism Speaks, and the Marino Autism Research Institute. My research efforts at the University of Southern California have been supported by a Dean's Fellowship (Viterbi School of Engineering, 2005-2009), the Simon Ramo Fellowship (Viterbi School of Engineering, 2009-2010), and the Alfred E. Mann Innovation in Engineering Doctoral Fellowship (2010-2012).

Table of Contents

List of Tables

List of Figures

Abstract

Human judgments on human behavior are an important part of interpersonal interactions and many assessment and intervention designs. While humans have evolved to be naturally adept at processing behavioral information, there are some challenges. Namely, human descriptions on behaviors are oftentimes qualitative, and there is variability between people's judgments due to the subjective nature of the judgment process.

Technology can help humans process behavioral data in a number of ways. Quantitative descriptors can be extracted from objective signals (e.g., audio, video) that represent aspects of human behavior in consistent and repeatable ways. There are many emerging engineering pursuits centered around modeling human behavior. Much of this research focuses on modeling specific human actions (e.g., head nods) during acted or non-spontaneous scenarios. Behavioral signal processing involves the development of computational methods that model human behavior in real-life scenarios. In this thesis, we automatically quantify and predict human subjective judgments on human behavior from speech signals in the context of societally-significant domain applications (education, family studies, health), where human observers play a critical role.

There are many technological challenges to quantifying and predicting human subjective judgments on human behavior. These include modeling several sources of variability,

including the human behavior itself (heterogeneity) and the human evaluators themselves. There is a need to extract robust generalizable features that capture the human behavior and the relevant perceptual cues human evaluators are using. In addition, there is possibly information across multiple modalities/cues, and it is not always clear how humans weight them when making their judgments. Many relevant human judgments are "gist-like," based off a large amount of behavioral data. Thus, modeling the data at possibly multiple granularities is important, since some temporal regions may be more relevant than others and a particular cue's importance may vary as a function of time. Finally, since we are analyzing real data in real-life scenarios, the human behavior can be complex and the data can be non-ideal (e.g., noisy).

For this thesis, we focused on concrete problem domains that highlighted specific aspects of the technological challenges: literacy assessment, couples therapy research, and autism diagnosis. In the literacy assessment domain, we show that we can exploit human-inspired information into the computational framework for accurate modeling of evaluator's perception of children's overall reading ability for one specific reading task. We fused features that represented multiples aspects of the human behavior and robustly emulated human observational subjective processes by learning from individual and multiple evaluator's judgments. We also exploit the fact that evaluators' level of agreement significantly varies (depending on the child being judged) by incorporating this source of evaluator variability in the modeling framework. In the couples therapy research, we analyze a large corpus of spontaneous dyadic interactions between married couples and show we can predict six relevant high-level observational judgments (e.g., level of acceptance, global negative affect) using speaker-dependent acoustic speech features. Furthermore,

we demonstrate one method for fusing automatically-derived speech and language information for improved classification of spouses' level of blame (high vs. low). Finally, we discuss our effort in collecting a multimodal corpus of child-psychologist interactions, recorded in the context of a social interaction used by psychologists for a research-level diagnosis of autism spectrum disorders. We highlight initial work with this corpus and discuss future experiments for the quantification of psychologists' clinical judgments on atypical social behavior (e.g., atypical prosody).

This thesis is on the development of a quantitative, automated framework that emulates human observational processes to describe human behavior from speech signals. We hope it makes impactful technological contributions to modeling complex human subjective processes. This work represents a significant step towards a shift in engineering from modeling and recognizing more objective human behaviors (e.g., speech recognition) to quantifying more subtle and abstract ones, a central theme to the emerging area of behavioral signal processing.

Chapter 1

Introduction

Understanding human behavior is a general goal of many fields in science. This includes understanding how people communicate, move, emote, and interact. There is also a need to understand the way in which people *judge* human behavior. Human judgments on human behavior occur everywhere: in everyday life (e.g., when judging the emotions displayed by a conversational partner), in educational settings (e.g., when teachers assess the reading skills of their students), in human-centered research (where oftentimes hypotheses are tested by manually coding relevant judgments on human behavior), and in clinical settings (e.g., when diagnosing psychological disorders).

Humans are naturally attuned to observe human behavior. For example, the human auditory system is optimized for perceiving human speech. However, there are challenges to human processing of behavioral information. First, humans lack the ability to quantitatively track or describe certain human behavior (e.g., the fine details of a speaker's pitch or the timing interplay between facial gestures). Second, there is variability due to the inherent *subjective* nature of some human judgments, which causes there to be differences between people's judgments due to many factors (e.g., background, expertise, mood). The next section is meant to illuminate these limitations of manually analyzing human behavioral data to motivate the use of automated computational methods.

1.1 Human Coding Complications

Oftentimes human behavior is recorded for off-line coding of relevant observational events; this is especially true when testing research hypotheses or for the purpose of training human coders. This manual coding is a costly and time consuming process. First, a detailed coding manual must be created, which often requires several design iterations. Then, multiple coders, each of whom has his/her own biases and limitations, must be trained in a consistent manner. The process is mentally straining for evaluators, and the resulting human agreement can be quite low. The following subsection provides a simple example illustrating one of the complications that can occur when manually annotating human behavioral data.

1.1.1 Example

The example we explain here is drawn from our experience when training two student evaluators to manually code specific social communication behaviors of children while interacting with a computer agent (work that is not included as part of this thesis [13, 14]). One of the behaviors we were interested in coding was when the child was smiling. The frequency of smiles (e.g., counts/minute) have been used in previous studies as a quantitative measure of shared enjoyment (e.g., in [173]). To ensure that the student evaluators could reliably code this behavior, we had them code instances of smiles on

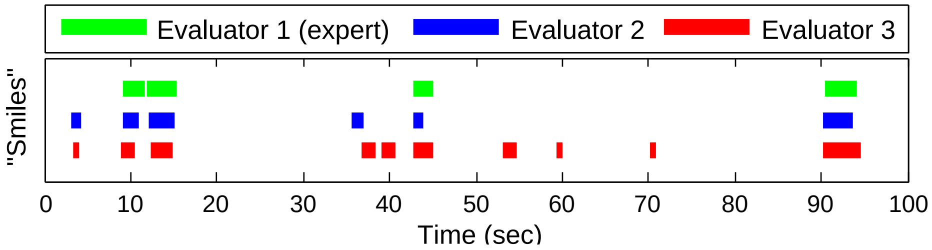

Figure 1.1: The manual coding results of 3 different evaluators (evaluator 1 is an expert psychologist, and evaluators 2 and 3 were trained students), coding a 100-second interaction between a computer character and a child. All three evaluators were shown the same video clip, and the three colored streams represent the time instances in which the evaluators recorded when the child was smiling. While there are temporal regions in which all three evaluators agree, there are noticeable coding differences between the three evaluators, which suggests that coding smiles is less objective than it may seem.

a few training videos. We also had an expert psychologist (who designed the coding manual) code these training videos.

Figure 1.1 shows the "smile" coding results of the three evaluators for one training video. It is clear from this figure that there are temporal regions in which all three evaluators agree. Yet, the labeling strategies of the three evaluators differ significantly, mainly on their strictness for what they consider a smile. That is, Evaluator 1 (the expert) marked fewer smiles than Evaluator 2, who in turn, marked fewer instances of smiles than Evaluator 3. This illustrates how subjective manual coding can be, even for seemingly simple human behavioral coding tasks. For reference, Figure 1.2 shows frames of the training video during regions in which 0, 1, 2, and 3 of the evaluators marked the frame as a smile.

This example motivates the need for a more objective way to code human behavior. One way to accomplish this is by using stringent coding systems for each relevant human behavior. For this smiling example, the Facial Action Coding System (FACS) provides

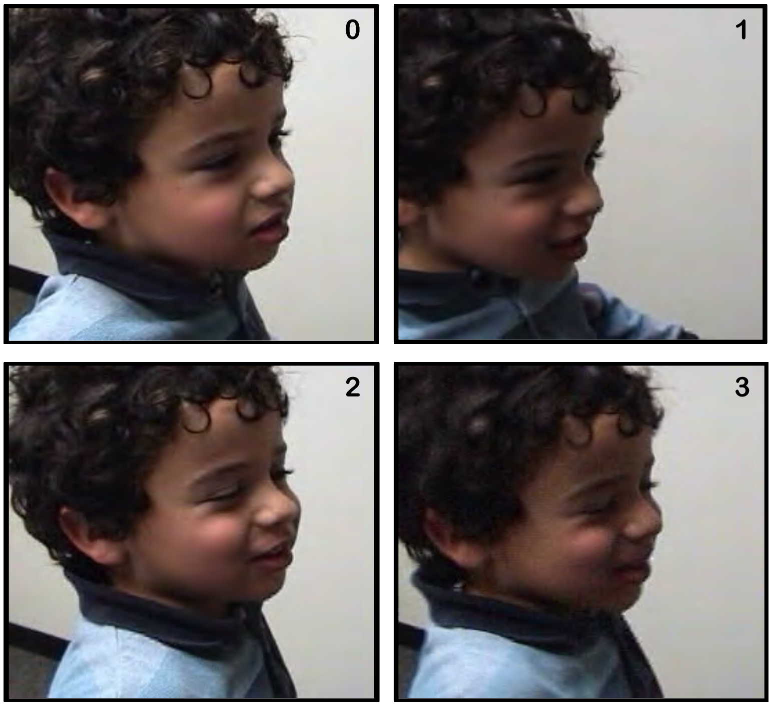

Figure 1.2: Selected frames during the 100-second child-computer interaction in which a variable number of the 3 evaluators recorded that the child was smiling. Frame 0 was a randomly selected frame in which none of the evaluators said the child was smiling. Only Evaluator 3 said the child was smiling in frame 1. Evaluators 2 and 3 marked frame 2 as the child smiling. All evaluators marked the child as smiling for frame 3.

detailed descriptions on a number of social human facial systems to help make this coding process more objective and consistent [68]. Even in this case, though, each coder needs to be trained to reliably code the data this way. Alternatively, one can imagine an automated smile detector that can track smiles in a continuous manner, thresholding the occurrence of a smile to match a particular coding style. This is currently being addressed in related research efforts [170, 176].

However, this thesis goes beyond modeling unimodal lower-level human behaviors such as smiling. It is primarily concerned with emulating *higher-level* subjective judgments, in which detecting smiles may be one useful cue/feature. For example, in [14], we showed that children playing a conversational, problem-solving computer game express *uncertainty* by using a combination of lexical (e.g., “I don’t know”), acoustic (e.g., question intonation), and visual/gestural (e.g., raised eyebrows) cues. In this work, we first had human evaluators mark which speaker turns/utterances they felt the child was uncertain. We then had evaluators manually code a number of potentially relevant audio-video cues. Finally, we used machine learning techniques to map the presence/absense of the cues within a turn to the perception of user uncertainty. In this dissertation, we examine these types of perceptually relevant learning problems, but critically, we find computational ways to model and predict them directly from the audio signal.

1.2 Thesis Statement

This thesis examines the modeling, quantification, and prediction of subjective judgments on human behavior in the context of societally-significant domains (education, family studies, health), where human behavioral evaluation plays a central role. This thesis addresses many of the technological challenges that have emerged as a result of working on this new class of problems, and the methods developed here can have broad implications in the new field we are calling *behavioral signal processing*, explained next.

1.3 Behavioral Signal Processing

Behavioral signal processing (BSP) is an emerging field in engineering. It encompasses the development of computational methods that model human behavior. This includes

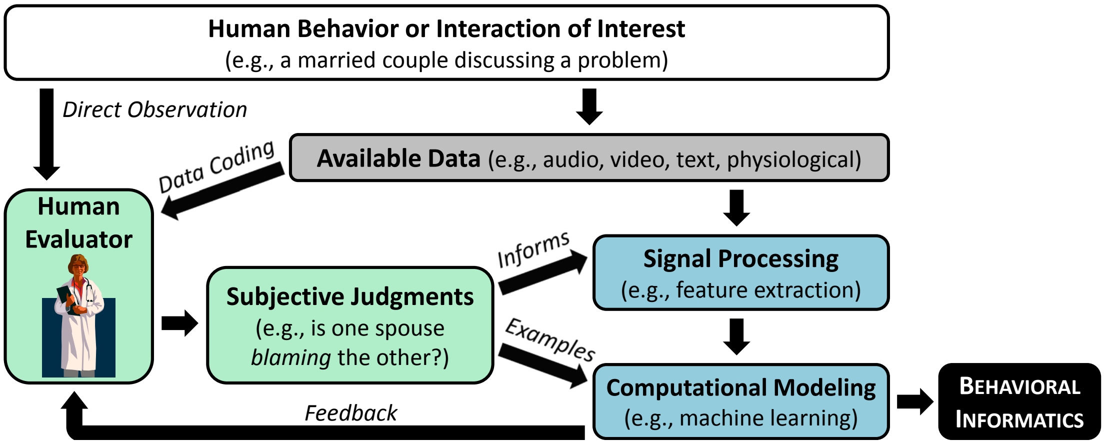
Human Behavior or Interaction of Interest
(e.g., a married couple discussing a problem)
Direct Observation
Available Data (e.g., audio, video, text, physiological)
Data Coding
Human Evaluator
Informs
Signal Processing
(e.g., feature extraction)
Subjective Judgments
(e.g., is one spouse blaming the other?)
Examples
Computational Modeling
(e.g., machine learning)
Behavioral Informatics
Feedback

Figure 1.3: A high-level flow chart that applies to a range of scenarios in behavioral signal processing (BSP) and shows the interplay between human evaluators and automatic computational algorithms.

emulating human-like observational and subjective processes. By *behavior*, we mean "anything that a person does involving action and response to stimulation [124]." There are multiple aspects of behavior that can be modeled, including those that are observable (and processed) by humans, and those that are manifested in physiological cues (such as autonomous responses to stimuli). While both can be built into the larger computational framework we call BSP, this thesis is primarily focused on the human observational and subjective processes. By *subjective*, we mean "modified or affected by personal views, experience, or background [125]." Therefore, a core aspect of BSP involves the development of objective technological tools that model two "human" aspects: 1) the conventionally-observable human behavior/interaction itself, and 2) the subjective judgments made by human evaluators who are observing the behavior/interaction. BSP can offer an invaluable ancillary to manual analysis in some cases, and can enable novel insights in others, for human-centered research and practice.

Figure 1.3 is a flow chart for a typical BSP scenario that shows the interplay between human evaluators and automatic computational algorithms. At the top of the chart, there is a human behavior or human interaction of interest (e.g., a married couple discussing a problem in their relationship). Human evaluators, who can range from trained coders to domain experts, are either directly observing the behavior/interaction or viewing the available data off-line and making relevant subjective judgments concerning the behavior/interaction (e.g., the "level of blame" expressed from one spouse to another in a conversation). This evaluation process can be formal or informal, explicit or implicit, expert-based or naïve-observer based. An example of a formal, explicit, expert-based evaluation would be a trained psychologist manually rating a child's social communication skills using a multi-dimensional coding scheme developed to help diagnose children with developmental disorders. An example of an informal, implicit, naïve-observer based evaluation would be a person sensing the emotions of a conversational partner in real-time.

On the computational side, signal processing methods transform the available data in some meaningful way, and computational modeling techniques (e.g., machine learning, estimation, fuzzy inference) map the signal cues to the human (often, fairly subjective) judgments of behavior. The human evaluator can aid in this automation by, for example, informing which signal features may be most relevant and also by providing labeled data used for automatic learning purposes. Conversely, the automatic output(s) can provide relevant feedback to the human evaluator(s) by automatically labeling new data and/or by offering some novel information and details about the human behavior/interaction ("behavioral informatics"). This back-and-forth information transfer can be computationally formalized and be iterated a number of times to form human-in-the-loop learning scenarios that refine the automatic algorithms to better complement or enhance human evaluations.

1.4 Technological Challenges

There are a number of technological challenges with quantifying and predicting subjective judgments on human behavior, due to the fact that modeling human behavior is inherently a complex problem. There is variability across both the human behavior itself (e.g., no two speakers speak in the same manner) and in the evaluators (due to the subjective nature of the judgments). The challenge of modeling this *production-perception* relationship is then two-fold: 1) there is a need to extract robust features from the available data that model the human behavior in a fashion that is generalizable within and across subjects, and 2) some processing of the subjective evaluations is oftentimes needed to allow for the computational methods to robustly learn human grading trends.

Furthermore, in regards to the extraction of useful features, it is not clear how human evaluators weight various modalities and behavioral cues when making their judgments. The extraction of features that completely cover the spectrum of cues that may be relevant to human evaluators is one of the critical challenges in BSP research, in addition to finding intelligent ways to merge, combine, or fuse multiple features. In addition, some temporal regions may be more relevant than others, and a particular cue's importance may vary as a function of time. One of the main challenges is the development of computational methods that are able to take into account the dynamic nature of the human behavior.

In addition, many human judgments are high-level or “gist-like,” based on a large amount of human behavioral data. These types of judgments are inherently fuzzy and qualitative by nature, since it is difficult for humans to pinpoint exactly how they are making them. Appropriately modeling the human behavior/interaction (at possibly multiple granularities) is important in order to emulate higher-level subjective judgments.

Finally, since we are analyzing human behavior in real-life scenarios, the data analysis is challenging. Human behaviors can be very complex (e.g., interpersonal dyadic interactions). In addition, human evaluators may have access to more information than is available to the computational algorithms (especially in cases in which the human evaluator was directly observing the human behavior/interaction). Cues that humans reliably use may not be able to be robustly extracted from the available signals. Also, there may be information loss (due to noisy environment conditions) or systematic data collection errors (such as poor microphone placement or inconsistent video angles across sessions). Appropriately dealing with adverse data conditions, while simultaneously ensuring *ecological validity* when collecting new data, are other aspects of BSP that are critical for the technology to be useful in real-life scenarios.

1.5 Three Application Domains

In this thesis, we focus on concrete problem domains to highlight specific aspects of the technological challenges described in the previous section. The empirical and applied aspects of this dissertation are developed based on data drawn from, and inspired by, societal problems in education, family studies, and health. Specifically, we will explore case studies that incorporate the central ideas of this thesis from three application domains: 1) children's literacy assessment, 2) couples therapy psychology research, and 3) autism diagnosis. These three domains involve societally significant problems that depend critically on human subjective judgments. Each domain addresses aspects of the general technological challenges, and we will explain the new computational contributions that have emerged as a result of tackling these unique problems. The following subsections give an overview of the three case studies highlighted in this dissertation.

1.5.1 Automatic literacy assessment

Literacy assessment is an important element in children's early education [28]. With respect to the BSP framework, the behavior of interest is children reading aloud, and the human evaluator is the teacher; assessments by the teacher include rating the children's correctness of pronunciation and judging the children's ability to fluently read at an acceptable rate. Education experts agree that one of the most effective assessment frameworks is *formative assessment*, where children are repeatedly assessed as they are taught [95]. Unfortunately, formative assessment is challenging for a number of reasons. First, assessment often requires one-on-one time, which teachers may not be able to provide, especially in large classrooms. Second, formative assessment requires an adaptive approach to teaching, where teachers are continually adjusting their lesson plans based on the children's rate of learning.

BSP can help with this process by emulating the teacher assessments and providing relevant feedback on the children's reading performance. In this thesis, the literacy assessment application is used to show that computational methods can mimic human grading patterns for “high-level” *overall* performance across a reading task. These assessments are subjective since overall judgments are less definable by nature, requiring evaluators to weight multiple aspects of the children’s reading to attain an overall grade. The children analyzed for this application of the thesis were from a diverse bilingual background and were recorded in actual kindergarten to second grade classrooms. There are a number of challenges in modeling this type of human behavior, due to the noisy environment conditions and the variability of children’s speech and second language learners [115].

Eleven human evaluators rated 42 children on their overall reading ability, after listening to recordings of them reading a list of English words aloud. We extracted multiple human-inspired features from the audio signal that were correlated with cues human evaluators stated they used: pronunciation correctness, speaking rate, and the fluency of the speech. Using linear regression techniques, we automatically predicted individual evaluators’ high-level scores with a mean Pearson correlation coefficient of 0.828, and we predicted average evaluator’s scores with correlation 0.952. These human-machine agreement statistics exceeded the mean inter-evaluator agreement, demonstrating the potential power in using features derived from objective signals. We also show the ability for automated methods to learn from multiple evaluator’s grading patterns, resulting in a robust automatic literacy assessment system that agrees with people’s perception significantly better, on average, than people agree amongst themselves.

1.5.2 Couples therapy research

Several fields in psychology depend critically on perceptual judgments made by people. Historically, behavioral psychology research has depended on manual analysis of human behavior [120], which can be a severe bottleneck in larger longitudinal studies. BSP can offer a powerful ancillary by providing a quantitative computational framework built on the processing of objective signals. In this thesis, we analyze a corpus consisting of real married couples spontaneously interacting about a problem in their relationship [45]. The coding manuals designed for this study had multiple trained evaluators rate each session with 33 high-level codes representing relevant aspects of *each* spouse's behavior (e.g., global positive affect expressed by the husband) [93, 99].

In initial experiments, our goal was to learn these high-level codes using audio features. This is a challenging learning problem due to the time scale of the evaluations; we were trying to predict session-level codes that represented the overall behavior of each spouse, a subtle "gist-like" judgment. In addition, since the data was originally only intended for manual coding, the recording conditions were not ideal for automatic analysis; the video angles, microphone placement, and background noise varied across sessions. By extracting a large number of speaker-normalized prosodic, spectral, and voice quality speech acoustic features, we were able to train models to classify extreme spouses' behavior (e.g., high vs. low level of blame) significantly better than chance for all six codes we examined. In addition, for the spouses' level of blame, we improved classification performance by incorporating important lexical blaming cues through fusion of automatically-derived speech and language information for multimodal prediction of evaluators' judgments.

1.5.3 Autism diagnosis

The final application domain in this thesis, which represents ongoing and future work, involves the diagnosis of autism spectrum disorders in children. Autism is a developmental disorder that results in impaired social communication and restricted, repetitive, and/or stereotyped behavioral patterns [3]. While studies have shown that an early diagnosis and an appropriate intervention can lead to improved communication skills in autistic children, diagnoses can be inaccurate, and interventions are oftentimes expensive and time-consuming [67, 149].

Our goal is to use computational methods to help psychologists and clinicians make the difficult judgments necessary when assessing children's social and communicative skills; this could be used for diagnostic purposes or to track children's progress during interventions. We are in the process of collecting a large multimodal (audio-video) corpus of children interacting with a trained psychologist in the context of the Autism Diagnostic Observation Schedule (ADOS) [118], a popular tool that psychologists use to help diagnose children with autism spectrum disorders. During the ADOS, the child interacts with the psychologist in a semi-constrained fashion that enables the psychologist to assess the child on a number of autism-relevant social and communication skills. One of the difficulties with the ADOS grading scheme is the qualitative nature for some of the codes.

The collection of this corpus is an important step in the development of technological tools for the automatic quantification of clinical judgments. Future plans involve the training of normative models of children's behavior and using data-driven methods with the corpus to automate aspects of the ADOS grading. Incorporating quantitative methods and models could lead to a more consistent grading scheme across subjects and over time. The technology could potentially be scalable to large populations of children.

1.6 Related Work

The work addressed in this thesis is inherently interdisciplinary, drawing from, and subsuming, problems being addressed in the behavioral sciences and borrowing from several related human-centered fields in engineering and computer science. There are a number of established or emerging disciplines of inquiry (often with intellectual overlap). These include *human behavioral analysis/understanding*, *social signal processing*, and *affect/emotion recognition*. We will introduce each field and discuss how they fit within the BSP umbrella.

Human behavioral analysis/understanding uses multimodal signal processing of auditory and visual information to model human actions in the context of human-human and human-computer interactions [1, 144]. This can range from the recognition of gestures (e.g., head movements [35,129]) to the modeling of spoken turn-taking behavior [112,128] to the detection of anomalous human activities (e.g., a person falling over [137]). Tracking, detecting, and categorizing human actions is an important element of BSP, since the behavior of interest in BSP research must be properly processed and analyzed.

Social signal processing attempts to understand the social signals expressed by people through the detection of lower-level human behavior, such as smiling and eye blinking [146, 174]. Similar to human behavior analysis/understanding, research in social signal processing overlaps with BSP topics since the detection of these lower-level signals is important when modeling human behavior. In addition, many of the societally relevant BSP topics involve a human social component, and accurately modeling these social signals becomes critical.

Affect/emotion recognition research involves the modeling and recognition of affective human behavior in realistic and acted scenarios using features derived from audio, video (including motion capture), language, and physiological signals [36, 86, 111, 126, 135, 153, 181]. This research naturally intersects with BSP topics, since affective states and emotions have both an expressive and perceptual component to them. In this regard, affect/emotion recognition can be considered a subset of the problems that BSP addresses.

BSP, along with the aforementioned related fields, represents a shift in engineering from modeling and recognizing more objective human processes (e.g., speech recognition) to quantifying more abstract ones. The hallmark of BSP, while it relies and leverages advances in the various related areas, is in extracting meaningful information about human behavior (*behavioral informatics*) in the context of societally significant application domains in which human evaluators play a critical role.

1.7 Contributions of the Thesis

In [25], we showed that disfluencies (e.g., hesitations, sound-outs, question intonations) in children's read speech were considered perceptually relevant to human evaluators when judging the overall reading ability of children. We devised a novel automatic speech recognition method that exploited the constraints of the reading task to automatically detect these speech disfluencies directly from the audio signal. In [23, 24], we extended this research to automatically predict evaluators' judgments on children's overall reading ability by extracting cues representing various aspects of the children's reading: pronunciation correctness, fluency, and speaking rate. Related work had concentrated on unimodal aspects of children's speech (pronunciation correctness or fluency or speaking rate), and fusing them together for one high-level prediction was a novel idea that enabled us to train models that more closely mimicked real reading evaluators like teachers. In [27], we provided an additional extension by showing that we can automatically learn both *individual* and *average* evaluators' grading trends. This body of work demonstrates how computational methods can learn from human evaluators at multiple stages: by informing feature extraction and by modeling multiple evaluators' perspectives.

In [17, 18], we provide details on experiments run on the couples therapy database. One of the main contributions of this thesis is the analysis of real data, and this corpus represents actual data recorded for a longitudinal study on the efficacy of a new form of couples therapy [45]. Manually transcribing and coding the data took a collaborative effort between multiple universities over a period of several years. Automatically analyzing real-life data collected from psychology-based studies is one novel aspect of this thesis. In [17, 18], we provide details on how we were able to align the text transcriptions to the audio signal for more than 60 hours of the couples' recordings. As part of our work, we showed we could separate extreme behaviors (as coded by trained evaluators) on relevant judgments (e.g., level of blame expressed by the husband) by extracting speaker-normalized speech features from the audio signal. In [16], we fused automatically-derived acoustic and language features for improved classification of spouses' level of blame. These technical and algorithmic contributions are very general and could be applied to any spontaneous interaction.

For the autism diagnosis research, the corpus we are currently collecting is an important novel contribution. Recruitment of subjects from protected populations (like children diagnosed with autism) is very difficult, which has led to smaller-scale studies that lack statistical significance and generalizability. We have already collected data from 70 subjects to date, which will allow for a larger-scale analysis of speaking trends across a heterogeneous subject population. As part of this thesis, we designed the audio-video recording set-up for a real clinic at Children's Hospital Los Angeles that records the interaction between the psychologist and child in a consistent and repeatable manner. Chapter 4 describes this USC CARE Corpus [26] and discusses future work we will carry out using this corpus.

This thesis provides in-depth analysis on the three aforementioned application domains that explores how computational models can help humans make subjective judgments on human behavior. The methods used in this paper are meant to be broadly useful for many problems that fall under the expanding BSP umbrella by addressing the general technological challenges in modeling high-level subjective judgments on realistic human behaviors.

1.8 Document Organization

As discussed earlier in the introduction, this dissertation is an analysis of specific case studies from three broad application domains that are but a part of the larger BSP puzzle. The rest of this dissertation is organized as follows. Chapter 2 examines work towards automatic literacy assessment. Chapter 3 discusses work on the couples therapy research database. Chapter 4 discusses ongoing and future intended research on how technology can help with autism diagnosis. Chapter 5 provides a conclusion and discusses open problems and future work.

Chapter 2

Automatic Literacy Assessment

2.1 Introduction

[1] Education is one area in which technology has already made a profound impact by providing an engaging learning experience to children [69]. Computer games have helped children develop problem-solving skills [181], and virtual peers have helped encourage creative thinking and children's use of imagination [40]. Literacy tutors have been developed to track children's reading and offer helpful feedback [89, 133]. These technologies have been designed for a range of ages and developmental levels, and for children with special needs [50].

While much research has focused on developing interactive educational technology, relatively fewer studies have concentrated on ways to use computer technology to help educators and teachers directly. We tackle this problem in the context of literacy assessment for young children from a diverse bilingual background that are learning to read English. Assessment of reading skills is an important aspect of early education [28]. Experts agree that one of the most effective assessment frameworks is *formative assessment*,

[1] This work was supported in part by the National Science Foundation.

in which teachers assess their students throughout the learning process. This pedagogical framework helps keep the teachers' goals and the children's progress aligned and prevents children from being left behind [95]. Unfortunately, formative assessment is challenging for a number of reasons. First, assessment is time-consuming since each child requires the full attention of the teacher. Second, formative assessment requires teachers to continually adjust their lesson plans.

Technology can help with this process in a number of ways. First, computers can be used to administer the various reading assessment tasks in a consistent, repeatable manner. Second, pronunciation verification systems can be developed to automatically assess the children's speech using objective signal-based methods (e.g., automatic speech recognition). And third, these results can be analyzed and displayed to teachers, so they can track the children's reading proficiency over time, and adjust their lesson plans accordingly.

Reading assessments can occur at different granularities (segmental or suprasegmental) depending on the intended application and reading task. For example, preliterate children are assessed on their knowledge of the letter-to-sound rules of a particular language, while more advanced students are assessed on their ability to fluently read phrases and sentences aloud [145]. Appropriate reading tasks must be designed to elicit speech that facilitates the intended assessment. One common theme among most reading assessment tasks is the use of multiple test items ("tokens") for each subject. This is done for a number of practical reasons. First, it ensures the subjects are provided enough tokens to cover many, or even possibly all, associated linguistic or category variations. Second, it allows evaluators to adjust to the speaking style of the subjects, so accent and idiosyncratic behaviors are taken into account. Third, it provides evaluators with statistically adequate evidence to make global ("high-level") assessments on the subjects' overall performance. We are specifically interested in this final aspect: to automatically model and predict evaluators' high-level assessments for a particular reading task widely administered to young children.

There is a need for technology that can be incorporated in the classroom to collaboratively assist in reading instruction [139]. We propose in this chapter to use automatic computer-based literacy assessments to help teachers, allowing them to better concentrate on lesson-planning and individualized teaching. Automatic computer-based literacy assessments can have several advantages over manual human-based assessments. Manual assessments are very time consuming, requiring one-on-one time. Doing continual assessments may not be feasible in a common scenario like a classroom, where there are several students and only one teacher, and where assessment time competes with instruction. Automatic assessment systems could significantly reduce the time burden of teachers. Manual assessments are also not standardized across evaluators, dependent on factors such as the evaluator's experience, personal biases, and human limitations (e.g., fatigue). Automatic computer-based assessments can provide a more consistent assessment framework, relying on objective features extracted from the available audio-video signals. A standardized computer-based automatic literacy assessment system could make more meaningful comparisons across children and over time. Finally, automatic literacy assessment systems can be portable and be scaled up to serve large populations of children.

There are several benefits for providing high-level overall assessments rather than (or in addition to) the more typical token-level assessments. First, having knowledge of the overall performance may be particularly useful when tracking performance over time. Second, high-level assessments provide a thumbnail view of a child's performance, which may be useful for teachers by aiding in instruction planning or designing further performance drill-down. Third, high-level assessments may model evaluators' perception better than token-level assessments. Whereas in token-level assessments, decisions are made on the goodness of that particular token, high-level assessments are directly modeling evaluators' interpretation on overall performance, which may be a multi-dimensional and/or non-linear mapping from token-level performance. Therefore, high-level assessments can be viewed as the interpretive extension to token-level assessments. Automatic high-level literacy assessment is a difficult problem because it involves the modeling and prediction of subjective human judgments. In order to accurately make high-level assessments, the multiple cues human evaluators might use have to be automatically extracted from the available measured observations. In addition, they have to be combined in a way that accurately models the high-level assessment. People might base their assessments on different cues when forming a grading criteria, and even in cases where evaluators use the same cues, they might differ on the relative importance of each. From a signal processing viewpoint, this requires the robust extraction of perceptually relevant features, followed by an appropriate machine learning algorithm that learns the interpretation of these cues, based on individual evaluators or a bank of evaluators.

There has been significant work on reading assessment, especially in second language learning and children's reading applications. Most of the related work has involved adults or children already reading phrases and sentences. We argue that literacy assessments at an earlier age is critical, since it has been shown that early literacy proficiency is a good predictor for reading fluency and comprehension proficiency in later grades [56,139,145]. Importantly, studies have shown a significant decrease in the percentage of poor readers when interventions take place before the second grade [143]. Automatic literacy assessments targeting younger children could help catch problems earlier, and an effective intervention could give children a better chance to grow into competent readers. In addition, much of the related work has concentrated on detecting segmental and suprasegmental errors in production for various reading tasks (e.g., [21,22,114,132,162–164,175]), but overall performance is rarely estimated. Some previous work has concentrated on providing overall scores (e.g., pronunciation quality [49], fluency [53], reading level [64]), but automatic high-level reading assessments remain relatively under-researched. It should be noted that the idea of modeling global holistic human judgments is not unique to literacy assessment. For example, the computer vision community has viewed this problem in the context of reconciling human evaluations and automatic scene classification [85,141]. Literacy assessments can fall under a number of overlapping reading-related skills, such as decoding words, fluently reading sentences aloud, reading comprehension, and writing. In this research, we assess children in kindergarten to second grade on their overall ability to fluently decode a list of English words aloud. This reading task is appropriate for this age group and resulted in speech that had a high level of variability in responses, including a range of disfluencies (e.g., hesitating, sounding out the words, elongating phones). While teachers can make use of both acoustic information and visual information (e.g., mouth movement, eye gaze) when assessing children's reading skills, we only have access to one audio signal, recorded from a close-talking microphone. Both the human evaluators and the automatic methods used this single audio channel, which may have resulted in a lower baseline performance for the human evaluators, as compared to a more traditional scoring setup. Future research will incorporate both acoustic and visual information to provide a more realistic scenario to human evaluators and to enable a multimodal approach to automatic literacy assessment. The combined use of audio and video information has been shown to bring increased accuracy and robustness in the context of automatic speech recognition [48, 117]. In this research, human evaluators listened to the children's speech and rated each on their overall reading ability on a Likert scale of 1 to 7. These human scores were the dependent variable for all our experiments and represented the high-level literacy assessment targets. There is always some level of subjectivity involved in assessment tasks, as is evident in variations across evaluators. Computers can help automate these types of judgments if they are able to make predictions that are in line with human evaluators. In this research, and in related research also involving human assessments (e.g., [86, 169, 179]), performance of the automatic system is measured by computing human-computer agreement. One could then view a computer as being competent if it can agree with human evaluators as much as humans agree amongst themselves. Ideally, computers would be able to adapt their grading styles to each evaluator or to a bank of evaluators.

In our previous paper [25], we showed that disfluencies have a perceptual impact on evaluators rating the overall performance of the children. We used a grammar-based automatic speech recognizer to detect disfluencies in the children's speech. In addition, we showed that by combining pronunciation correctness, disfluency features, and temporal speaking rate features, we could predict the average evaluator's scores with agreement that was comparable to human inter-evaluator agreement [23, 24]. We improve upon our pronunciation verification and disfluency detection methods and train a system using various feature selection procedures and linear regression techniques. We also extend our analysis to predict individual evaluator's scores. The final optimized system was able to learn both an individual evaluator's high-level scores and the average evaluators' scores with a higher level of agreement with which evaluators agree among themselves [27].

This chapter is organized as follows. Section 2.1 discusses the TBALL Project and the TBALL Corpus, on which this work is based upon. Section 2.2 describes and analyzes the human evaluations we administered to attain perceptual judgments. Section 2.3 discusses the features we extracted that are correlated with the cues evaluators used when making high-level judgments. Section 2.4 discusses the general machine learning methods we studied to predict evaluators' high-level assessments, and Section 2.5 provides our results and discussion. We propose one additional machine learning method that incorporates evaluator variability/uncertainty in Section 2.6. Finally, we provide a discussion in Section 2.7.

2.2 TBALL Project and Corpus

The Technology-Based Assessment of Language and Literacy (TBALL) Project was formed to create automatic literacy assessment technology for young children in early education from multi-lingual backgrounds [2, 147]. The TBALL Project's main goal was not to create real-time automated literacy tutors (see [50, 65, 88, 90, 131, 133, 177] for examples) but rather to provide a technological assessment framework that teachers could use to inform their teaching and track children's progress on age/grade-specific reading tasks. The reading tasks were designed for and administered to children in actual kindergarten to second grade classrooms in Northern and Southern California. About half of the children were native speakers of American English, with the other half non-native or bilingual speakers of English from a Mexican-Spanish linguistic background. The young age of the children and diverse population make this project and resulting corpus unique from other existing corpora [7, 70, 159].

We administered different reading tasks, compared to other automatic literacy assessment projects, to be more geared to preliterate children. These ranged from testing the production of English letter-names, the sounds corresponding to each letter ("letter-sounds"), syllable-blending tasks, to reading a list of isolated words [19, 21, 22]. The resulting speech from a single close-talking headset microphone makes up the TBALL Corpus [106]. Since the reading tests were administered in actual classrooms, the background noises included typical classroom sounds, such as other children's voices and the teacher's voice. The children's demographics (gender, grade, native language) were obtained by forms filled out by assenting parents and were included as part of the corpus when available.

For this work, we analyzed speech from an adaptation of the Beginning Phonic Skills Test (BPST), an isolated word-reading task consisting of 55 predetermined words. This word list was chosen since it evaluates children's phonemic awareness and decoding skills [56]. The difficulty of the words is steadily increased throughout the reading task, starting with monosyllabic words (e.g., map, left, cute), and ending with multisyllabic words (e.g., silent, respectfully). When administering the test, each word was displayed on a computer monitor one at a time, and the children had up to five seconds to say the word aloud before the next word was shown. The children had the option to advance to the next word before this five-second limit by pressing a button. During the data collection process, a trained research assistant listened beside the child, and if the child mispronounced three words in a row, the assistant manually stopped the session. This was done to prevent the children from getting too frustrated and is not the termination criterion from the BPST as generally administered. As a result, only 11.0% of the children read the full list of 55 words from our sample (M = 21.6 words, SD = 11.2 words). The transition times between words were automatically recorded, and these times were used to split each child's audio into single-word utterances.

Our test set was comprised of the speech from 42 children, each of whom completed at least the first ten words of the isolated word-reading task. These children were selected from a total of 100 children's data to ensure a wide variety of performance levels and reading styles and to be near balanced with respect to gender and native language. We chose 42 children to limit the total amount of speech to approximately 30 minutes to prevent evaluator fatigue when manually assessing the speech (described in Section 2.2). To ensure the words read by each child were of comparable difficulty, we only selected words that appeared in the top 25 of the word list. In total, the test set had 770 single-word utterances, an average of 18.3 words per child (SD = 5.07 words). The final demographics of the 42 children were: gender (female=21, male=21), grade (kindergarten=5, first=22, second=15), and native language (English=20, Spanish=18, bilingual=4). We also constructed a held-out feature development set with 220 children's speech from the isolated word-reading task; this set is described in detail in Section 2.3.2. Lastly, we used 19 hours of held-out speech from word-reading and picture-naming tasks to train 33 monophone acoustic models, a word-level filler "garbage" acoustic model on all speech segments, and a background/silence acoustic model on background segments of the recordings. All acoustic models were three-state Hidden Markov Models (HMMs) with 16 Gaussian mixtures per state. For features, we extracted a 39-dimensional vector, consisting of the first 12 Mel-Frequency Cepstral Coefficients (MFCCs), log energy, and their delta and delta-delta coefficients, every 10 ms using a 25 ms Hamming window. We applied cepstral-mean subtraction across each single-word utterance to help make the features more robust to classroom noise. We used the Hidden Markov Model Toolkit (HTK) [183] for all MFCC feature extraction, acoustic model training, and decoding.

2.3 Human Evaluations

2.3.1 Evaluation 1: High-level Literacy Assessment

Evaluation 1 was administered to obtain human perceptual judgments of high-level literacy assessments for the 42 children in the test data. Eleven English-speaking volunteers rated the children on their "overall reading ability." The evaluators fit into four classes: three had worked on children's literacy research for over a year, three were linguists, four were non-native speakers of American English with an engineering background in speech-related research, and three were native English-speaking individuals with no linguistics background or experience with speech or literacy research; the evaluators belonged to only one of the four classes, except for one linguist who also worked with children's speech and a different linguist who was a non-native speaker. While none of the evaluators were licensed teachers or reading experts, we found in previous work that the inter-evaluator agreement between teachers and non-experts was not significant for a pronunciation verification task [165]. Analysis of the inter-evaluator agreement for the 11 evaluators in this work will be provided in Section 2.4. The order of the children was randomized for each evaluator, but the word order within each child's session was maintained. The evaluators were provided the word list, so they could follow the children's progress. A short beeping sound was inserted between each single-word utterance, so the evaluators knew when the transitions between words took place. After listening to the speech from a child, evaluators rated her/his overall reading performance on an integer scale from 1 ("poor") to 7 ("excellent"). Examples of a "poor" reader versus an "excellent" reader were not provided to the evaluators beforehand for two reasons: 1) we did not know in advance whether all evaluators would agree on what a "poor" versus an "excellent" reader was, and 2) we wanted evaluators to come up with their own grading criteria for this reading task. Since evaluators likely needed to listen to a few children before getting comfortable with their own grading scheme, they were permitted to change previously assigned scores.

After the evaluators rated the 42 children, we asked one open-ended question to find which criteria evaluators used when grading the children. This was done to get a rough estimate of the relative importance of various cues people used for this assessment task. The evaluators' responses were grouped into three categories: pronunciation correctness (stated by 10 out of the 11 evaluators), fluency (stated by 9 of 11 evaluators), and speaking rate (stated by 9 of 11 evaluators). It should be noted that none of the evaluators specified that they based their judgment on the child's relative performance at the beginning or end of the word list or on the number of words spoken by the child. The number of spoken words was somewhat artificial for this data, since a human evaluator will not be present to stop the session if the task were administered by a computer; therefore, we do not use the number of words the child spoke as a feature for automatic high-level literacy assessment. While word order and word difficulty most likely had some effect on human evaluators, we assumed each word was equally important. Coming up with a quantitative system that takes into account a word's importance based on its location in the word list is difficult because these effects are most likely evaluator-dependent. The fact that children read a variable number of words from the word list further complicates the matter. Future work could use machine learning algorithms that take into account word list effects by weighting words differently, as was done in our previous work [162].

Based on the evaluators' responses, we concentrated on automatically extracting features/scores from the audio signal that correlated with pronunciation correctness, fluency, and speaking rate. There has been a significant amount of research on automatic pronunciation verification (accepting or rejecting the pronunciation of a target word), and we will employ some of these techniques on the development set in Section 2.3.3. Speaking rate features and other temporal correlates are also straight-forward to extract if the word pronunciations can be correctly endpointed. However, quantifying fluency is a more difficult task, since we did not know what made a response "fluent." To discover this, we used a second human evaluation, described next.

2.3.2 Evaluation 2: Perceptual Impact of Disfluencies

Evaluation 2 explored the impact of fluency on people's perception. We noted five main "disfluencies" in the data: hesitations, sound-outs, elongations of phones, whispering, and speaking with a questioning intonation (perhaps expressing uncertainty). Here, we use the term "disfluency" to describe any speech phenomena that takes away from the natural flow of the pronunciation of the target word. Typically, the term disfluency is used in the context of spontaneous speech for events like fillers (e.g., "uh"), repetitions, repairs, and false starts [161]. However, since this is a reading task and the children are learning how to read (and some are still learning how to speak English as a second language), the types of disfluencies are different from those studied in adult spontaneous speech.

We prescribed a set of conditions necessary for each disfluency type to make the task of labeling disfluencies more objective. The types of disfluencies that occurred in the data before the target word pronunciation included hesitations, where the child started to pronounce the target word, paused, and then said the target word, and sound-outs, where the child pronounced each phone in the word, pausing between each one, and then pronounced the target word. Some children whispered when sounding-out and hesitating, speaking voiced phones in an unvoiced manner. The other two types of disfluencies we noted took place during the pronunciation of the target word. Some children lengthened a phone or syllable of the target word, which we call elongations. Lastly, some children's pitch rose at the end of a word's pronunciation, which we refer to as a question intonation. It should be noted that these disfluency types were not mutually exclusive within an utterance. For example, a child might hesitate at first, and then say the word with a question intonation, or a child might use a whispered voice while sounding out the word.

For this evaluation, we selected 13 children's speech from the test set which displayed varying levels of the five disfluency types. Since labeling disfluencies is partially subjective, we had two evaluators (the first and second authors) mark each utterance with the presence/absence of each disfluency type. Table 2.1 shows that the percent agreement between the evaluators was high, so we used Evaluator 1's labels as the ground-truth for the remainder of our analysis. We then had 16 evaluators (eight engineers with speech-related background, four with teaching experience, and four with a linguistics education) rate for each word utterance the fluency of the speech (on an integer scale from 1 to 5). The words were grouped by child, so evaluators could adjust to the speaking style of the children. The resulting fluency scores from the multiple evaluators were transformed to z-scores by subtracting the mean of each evaluator's scores and dividing by the standard deviation. This normalization was done to allow for more meaningful comparisons of scores between evaluators. We found that the mean normalized fluency score for utterances that contained no disfluencies (M = 0.637, SD = 0.792) was significantly higher than the mean score for utterances that contained at least one disfluency type (M = -0.484, SD = 0.854), $t(2035) = 30.3$, $p < .001$. This shows that indeed utterances which were not labeled with any of the five disfluency types were considered more fluent. We also computed pairwise one-sided t-tests to compare the mean normalized fluency scores between disfluency types. Table 2.2 shows that the sound-out and hesitation disfluencies were considered the most disfluent, and utterances with whispers were considered more disfluent than ones with question intonations or elongations.

To discover the relative contribution of each disfluency type on the perception of fluency, we also ran a regression analysis. The dependent variable was the vector of normalized fluency scores, and the independent variables were the binary ground-truth labels of the five disfluency types for each utterance. We found these independent variables were able to account for a significant portion of the variance in the fluency scores, $R^2 = .331$,

Disfluency	Frequency Counts (out of 146)		% Agreement
	Evaluator 1	Evaluator 2	
Sound-out	39	38	97.95
Hesitation	27	29	97.26
Whisper	22	26	97.26
Elongation	13	22	93.84
Question	10	14	95.89

Table 2.1: The number of utterances (out of 146) that each evaluator labeled as containing each of the five disfluency types and the percentage of utterances in which the two evaluators agreed.

Disfluency	M	SD	p-value			
			Hes	Wh	Qu	El
Sound-out	-0.648	0.865	0.154	0.001	$< .001$	$< .001$
Hesitation	-0.587	0.804	–	0.015	$< .001$	$< .001$
Whisper	-0.397	0.946	–	–	0.011	0.012
Question	-0.210	0.714	–	–	–	0.271
Elongation	-0.164	0.672	–	–	–	–

Table 2.2: Statistics of the normalized fluency scores for each of the five disfluency types, along with the resulting p-values when using pairwise one-sided t-tests to compare the difference in mean scores.

$F(5,\ 2031) = 201.0$, $p< .001$. As shown in Table 2.3, the coefficient magnitudes for the sound-out, hesitation, and whisper disfluencies were largest, which suggests their presence impacts evaluators' perception of fluency more than the elongation and question intonation disfluencies.

We conjecture that whispers, hesitations, and sound-outs were considered more disfluent because they occurred in addition to the pronunciation of the target word, thus breaking up the flow of the speech more than disfluencies that occurred during the pronunciation of the target word. Based on these results, we set out to automatically detect these three perceptually relevant disfluencies directly from the audio signal. Section 2.3.4

Disfluency	Coefficient	Std. Error	$t(2031)$	p-value
Sound-out	-1.206	0.045	-26.68	$< .001$
Hesitation	-1.047	0.052	-19.99	$< .001$
Whisper	-0.718	0.078	-9.224	$< .001$
Elongation	-0.500	0.072	-6.930	$< .001$
Question	0.150	0.057	2.645	0.008

Table 2.3: Regression analysis of the five disfluency independent variables when estimating the evaluators' normalized fluency scores.

discusses our proposed methods and shows results based on experiments with the development set.

2.4 Feature Extraction

We learned in Evaluation 1 (Section 2.2.1) that people considered pronunciation correctness, fluency, and speaking rate to be critical cues in determining the child's overall reading ability. In Evaluation 2 (Section 2.2.2), we learned that the whispering, hesitation, and sound-out disfluencies were considered the most perceptually relevant. In this section, we concentrated on extracting features correlated with these cues. In Section 2.3.1, we describe the construction of a dictionary for each target word, which we will use for much of our subsequent analyses. In Section 2.3.2, we describe the development set in greater detail. In Section 2.3.3 and 2.3.4, we use this development set to experiment with automatic pronunciation verification and disfluency detection methods, respectively. In Section 2.3.5, we apply these methods to the test data to extract features for high-level literacy assessment.

2.4.1 Dictionary

For each target word, we constructed a dictionary with the help of an expert teacher and linguist. Acceptable and foreseeable unacceptable phonemic pronunciations were included in each target word's dictionary. These unacceptable pronunciations were made by substituting correct pronunciations with common letter-to-sound errors; for example, /k ah t/ ("cut") was augmented to the dictionary as a common reading mistake for /k y uw t/ ("cute"). Also, due to the large Mexican-American background in the corpus, we added common Spanish-speaking influenced variants to the dictionary, based on [182]. On average, each target word had 1.20 acceptable pronunciations and 3.03 foreseeable unacceptable pronunciations in its dictionary. Across all target words, 33 phonemes were used in these pronunciations. (We trained a monophone HMM for each, as described in Section 2.1).

2.4.2 Feature Development Set

To test various feature extraction methods, we used the development set, introduced in Section 2.1; this speech data was not included in either the test set or the acoustic model training data. Most of the demographic information about the 220 children was unknown, since the children's parents did not provide this optional information: gender (female=25, male=43, unknown=152), grade (kindergarten=5, first=36, second=27, unknown=152), and native language (English=21, Spanish=38, bilingual=5, unknown=156).

Since we were interested in detecting mispronunciations and disfluencies as relevant features, we first needed to explicitly label these in the development set. Three evaluators manually verified the pronunciation of each target word in the development set (binary accept/reject) and labeled each single-word utterance with the five disfluency types. All utterances in which there was excessive background noise or problems during the recording (e.g., cut-off speech) were marked by the evaluators and ignored. There was no overlap in evaluations, since this manual labeling process is costly (we saved approximately 20 hours of time by using three evaluators with no overlap) . In total, 2800 single-word utterances were annotated. 22.95% of the utterances had at least one disfluency type, and 2.49% had two or more types. Hesitations were marked in 8.93% of the utterances, sound-outs in 5.94%, elongations in 5.15%, whispering in 3.13%, and question intonations in 2.13%. 37.1% of the target word pronunciations were rejected. If at least one disfluency was marked in the utterance, the probability the pronunciation was rejected increased to 0.578. This means that disfluent speech and mispronunciations were positively correlated events.

2.4.3 Automatic Pronunciation Verification

The purpose of automatic pronunciation verification is to accept or reject a pronunciation. To characterize the performance of this task, we borrow metrics commonly used in detection theory and binary classification tasks: precision (2.1), recall (2.2), balanced F-score (2.3), false-alarm rate (2.4), misdetection rate (2.5), and Matthews correlation coefficient (2.6). In these equations, a true positive (TP) is correctly detecting a mispronunciation, a false positive (FP) is incorrectly detecting a mispronunciation, a true negative (TN) is correctly detecting no mispronunciation, and a false negative (FN) is incorrectly detecting no mispronunciation.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.2}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{2.3}$$

$$FA = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{2.4}$$

$$MD = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}} \tag{2.5}$$

$$MCC = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})} \tag{2.6}$$

In our previous papers [23, 24], we used a simple automatic pronunciation verification method, which acts as our baseline method for this work. We ran automatic speech recognition (ASR) with the dictionary of acceptable and foreseeable unacceptable pronunciations on each single-word utterance in the development set. We tried a number of different finite-state grammars (FSGs) to endpoint the pronunciation automatically: allowing for recognition of the background model (BG) vs. the garbage model (GG) at the start and end of the utterance vs. allowing both to be recognized; requiring the BG or GG models to be recognized at the start and end of the utterance vs. making it optional; allowing for repetitions of the BG and GG models at the start and end of the utterance vs. only allowing them to be recognized once. We found, in general, that allowing for the GG model to be recognized at the start and end of the utterance resulted in more false alignments of the target word pronunciation, probably because the GG model was trained on speech data. Fig. 2.1 shows an example of the FSG that attained the highest

F-score. In this FSG, the BG model is recognized (with the option of multiple recognitions) at the start and end of each utterance, and there is one required forced alignment of either the background model (BG), the garbage model (GG), or one of the acceptable or unacceptable pronunciations in the dictionary for that target word. A pronunciation is accepted if and only if an acceptable pronunciation of the target word is recognized; otherwise, it is rejected. The first row of Table 2.4 shows the performance of this method (called LEX), with respect to the metrics (2.1)-(2.6).

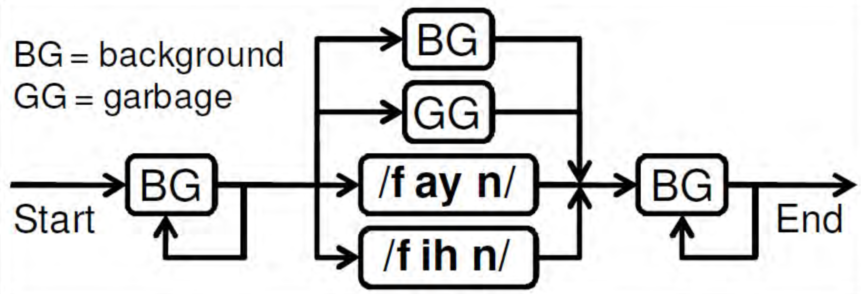

Figure 2.1: The finite-state grammar (FSG) used for the LEX pronunciation verification method (for the sample word, "fine"). The pronunciation is accepted if and only if the correct pronunciation (/f ay n/) is recognized; otherwise, it is rejected.

The second automatic pronunciation verification method we tried was Goodness of Pronunciation (GOP) scoring [179]. In this method, a forced alignment of acceptable pronunciation(s) of the target word is first made to the utterance. The resulting output will contain the phonemes recognized and their corresponding boundaries and acoustic log-probabilities. An unconstrained phone loop is then decoded across each phone segment, and a final GOP score for each phone is computed by subtracting the acoustic log-probability of the phone loop from the log-probability of the forced-aligned phone. High GOP scores correspond to phones that are more likely to be correctly pronounced, and a GOP score threshold can be made to reject phones with GOP scores below the threshold.

We applied this technique to each utterance in the development set and got the best results, in terms of maximizing F-score, when we did not threshold on individual phones within a target word but rather thresholded on the average GOP score across the word (where each phone is counted equally). Equation (2.7) shows how to compute the GOP phone score (O is the acoustics, p is the phone, PL is the phone-loop, and N is the number of frames of phone p). Equation (2.8) shows how to compute the GOP word-level score, by calculating the mean of the GOP phone scores for the word. Finally, (2.9) shows how we thresholded the GOP word-level score to ultimately reject or accept the pronunciation. This threshold, T, can be chosen to attain specific performance characteristics; we chose the T that maximized F-score, but other popular optimization criteria could be used (e.g., equal precision and recall, equal false-alarm and misdetection rates, maximum Matthews correlation coefficient). Table 2.4 shows the performance of this GOP scoring method for this optimal value of T.

$$\mathrm{GOP}(p) \equiv \frac{1}{N} \log \frac{P\left(O|p\right)}{P\left(O|PL\right)} \tag{2.7}$$

$$\mathrm{GOP}(l) \equiv \frac{1}{|p \in l|} \sum_{p \in l} \mathrm{GOP}(p) \tag{2.8}$$

$$\mathrm{Reject}(l) \equiv \begin{cases} 1, & \mathrm{GOP}(l) \leq T_{\mathrm{GOP}}(l) \\ 0, & \mathrm{GOP}(l) > T_{\mathrm{GOP}}(l) \end{cases} \tag{2.9}$$

We also tried combining the LEX and GOP methods. The LEX method makes use of target word knowledge and common letter-to-sound mistakes a child might make (especially with the influences of Spanish), but this method may be unable to detect errors if the child produces an unforeseeable realization of the target word. On the other hand, the GOP method is able to detect errors made that were not foreseeable but might not be able to tease apart close pronunciations with one phone substitution. We combined the two methods by first running the LEX method and then using the GOP scoring method only on pronunciations that were accepted by the LEX method. Table 2.4 shows results for all three proposed pronunciation verification methods, and Figs. 2.2 and 2.3 show performance as a function of GOP score threshold. We attained the highest F-score (0.802) and Matthews correlation coefficient (0.680) by using the combined LEX + GOP scoring method.

System Type	R	P	F	MD	FA	MCC
LEX	0.702	0.826	0.759	0.298	0.087	0.639
GOP	0.785	0.785	0.785	0.216	0.127	0.657
LEX+GOP	0.832	0.775	0.802	0.168	0.143	0.680

Table 2.4: Performance of the pronunciation verification methods: LEX, GOP, and the combination LEX+GOP, in terms of (2.1)-(2.6). The LEX+GOP method attained the highest F-score and MCC.

2.4.4 Automatic Disfluency Detection

Since this is a reading assessment task, the target words are known ahead of time. Furthermore, the sounding-out, hesitation, and whispering disfluencies were partial word manifestations of some pronunciation variant of the current target word. This facilitated the use of automatic speech recognition using finite-state grammars (FSGs) to detect

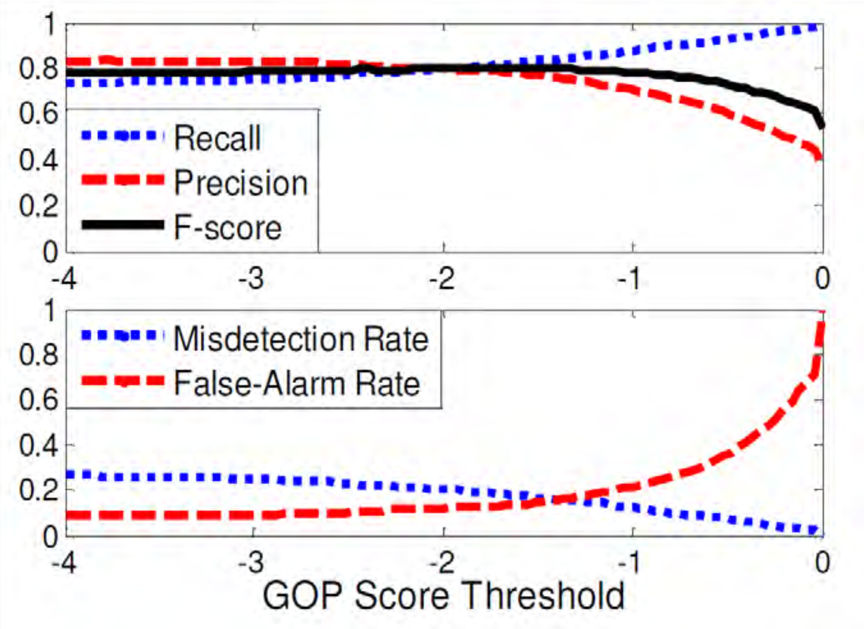

Figure 2.2: Performance of LEX+GOP pronunciation verification method as a function of the GOP score threshold (all pronunciations with GOP scores lower than this threshold were rejected).

disfluent speech. We first developed two simple baseline FSGs. The first baseline (Base1) allowed for repetitions of the target word with optional silence decoded in between. If two or more target words were recognized, the utterance was deemed disfluent; otherwise, it was deemed fluent. This baseline was chosen since the disfluencies usually consisted of phonemes that were present in the target word. The second baseline (Base2) inserted a phone loop (again with optional silence decoded between phones) prior to a required forced alignment of the target word. If one or more phones were recognized, the utterance was deemed disfluent; otherwise, it was deemed fluent. This second baseline was chosen since oftentimes the full target word was not spoken during a disfluency, so a phone loop allowed for partial words to be recognized. Table 2.5 shows the performance of these two baselines, in terms of the same six metrics we used before (2.1)-(2.6). Here, a "true positive" is the correct detection of a disfluency. As shown in Table 2.5, Base1 suffered from low recall (high misdetection rate), since the grammar was unable to recognize partial words, while Base2 suffered from low precision (high false-alarm rate), since its unconstrained phone loop resulted in a high number of false alarms.

To improve upon these baselines, we created a two-stage procedure for detecting disfluencies that combined both baselines, allowing for partial words to be recognized using only phones present in the target word. In the first stage, we designed a disfluency-specialized FSG to ensure a low misdetection rate (high recall). In the second stage, we rejected some of these detections to reduce the false-alarm rate. The first stage in the disfluency detection was introduced in [25] and based on work in [87, 89, 91]. We created target-word specific FSGs to recognize partial words. Since most disfluencies were partial word manifestations of the target word (or a partial word manifestation of a common mispronunciation of the target word), we created constrained FSGs that only allowed phones in the target word to be recognized and only in the order they appear in the dictionary. We experimented with many FSG designs: an unconstrained phone-loop consisting only of phones within the target word pronunciation(s) vs. requiring phones to be recognized in the order they appear in the target word pronunciation(s); allowing for repetitions and skipping of phones; requiring the first phone to be recognized vs. allowing it to be skipped; and allowing for optional repetitions of the BG model to be recognized between phones. All the FSG designs had high recall statistics above 0.94, so we chose to use the FSG shown in Fig. 2.4, since it had the highest precision statistic (Table 2.5).

Analyzing the errors made in stage 1, we noticed that many of the false-alarms were due to the recognition of unvoiced phones like stops (/k/, /p/) and fricatives (/f/, /s/).

These “noise-like” phones were similar to the classroom noise, and therefore, more susceptible to false alarms than vowels and other voiced phones. We tried a number of methods to reject some of these false alarms while still maintaining a low misdetection rate: 1) rejecting utterances below a minimum number of partial words recognized, 2) rejecting partial words that were below a minimum length in time, 3) rejecting partial words that were below a minimum acoustic model log-likelihood, 4) rejecting partial words that were below a minimum GOP phone-level score (2.7). We got the best results, in terms of maximizing F-score, by rejecting recognized partial words that were shorter than a minimum time threshold. Figs. 2.5 and 2.6 show how these performance metrics vary as a function of the threshold, and Table 2.5 shows the performance of the proposed two-stage disfluency detector when using the threshold that maximized F-score.

Compared with the two baseline methods, we attained the highest F-score (0.783) and Matthews correlation coefficient (0.737) with this two-stage FSG method. Further examining the performance of the two-stage FSG method when choosing the threshold that maximizes the F-score, 94.35% of the hesitations and 93.94% of the sound-outs were successfully detected. It most likely was unable to detect as many instances of whispering (58.62%) because of acoustic mismatches with the non-disfluent speech we used to train the acoustic models. In addition, whispered speech is more likely to be dominated by background noise.

2.4.5 Feature Extraction on the Test Data

We next applied these pronunciation verification and disfluency detection methods on the test data to extract scores correlated with evaluators’ perception of the children’s

System Type	R	P	F	MD	FA	MCC
Base1: Word Reps	0.175	0.965	0.297	0.825	0.001	0.376
Base2: Phone Loop	0.989	0.273	0.428	0.011	0.568	0.336
FSG: Stage 1	0.942	0.611	0.741	0.058	0.129	0.697
FSG: Stage 2	0.885	0.702	0.783	0.115	0.081	0.737

Table 2.5: Performance of the disfluency detection methods: baseline 1 (Base1), baseline 2 (Base2), and the 2 stages of the target word-specific finite-state grammar (FSG) procedure. The proposed 2-stage FSG method achieved the highest F-score and MCC.

reading ability. Since this was an isolated word-reading task, we extracted all features at the word-level. Table 2.6 shows the 48 scores extracted for each word. There are 10 scores based on the pronunciation verification methods, 12 scores based on the disfluency detection methods, and 26 speaking rate and other temporal scores based on both methods. When applying the pronunciation verification and disfluency detection methods discussed in Section 2.3.3 and 2.3.4, we used all threshold and parameter values that maximized the F-score on the development set. Note that we extracted the square root of all temporal features as an additional feature. This was done since the temporal features oftentimes had distributions that were skewed because of a small percentage of long times. The square root helped push the distributions towards a more bell-shaped distribution, which better fit the distributions assumed in the linear models we applied in Section 2.4. We found this square root transformation performed empirically well in our previous work [24]; future work could find a more optimal transform by choosing the root that makes the distribution most "normal." We extracted our final set of features for each child by computing 12 statistics across each word-level score for all the words spoken by the child: mean, standard deviation, skewness, minimum, minimum location (normalized by number of words spoken by child), maximum, maximum location (normalized), range, lower quartile, median, upper quartile, interquartile range. This produced our final feature set of 576 features per child. The next section will discuss how we used feature selection and supervised learning algorithms to properly deal with this over-generation of potentially useful features.

2.5 Prediction of Children's Reading Ability

Section 2.3 explained our feature extraction, which resulted in 576 child-level features. In this section, we used this feature set to predict children's reading ability, as rated by the 11 evaluators (see Section 2.2.1). Since there were 11 evaluators, there were many ways to pose this learning problem. We first analyzed the inter-evaluator agreement of the evaluators using Pearson's correlation coefficient. Equation (2.10) is Pearson's correlation between two vectors of scores, y_1 and y_2, where $y_j = \left[(y_j^1 \ldots y_j^{42}\right]^T$, and μ_{y_j} is the mean score for y_j. Note that the "42" in this equation refers to the total number of children we are assessing.

$$Corr(y_1, y_2) \equiv \frac{\sum_{i=1}^{42} (y_1^i - \mu_{y_1})(y_2^i - \mu_{y_2})}{\sqrt{\sum_{i=1}^{42} (y_1^i - \mu_{y_1})^2 \sum_{i=1}^{42} (y_2^i - \mu_{y_2})^2}} \tag{2.10}$$

Table 2.7 shows the pairwise inter-evaluator agreement using (2.10) and also displays four sets of average agreement for each evaluator. All 11 evaluators' scores had higher correlations with ground-truth scores (computed by averaging the other evaluators' scores), as compared to the mean pairwise correlation with the other evaluators. This means that the ground-truth scores are representative of the "average" evaluators' perception. In addition, for 9 of the 11 evaluators, agreement was higher when using all evaluators to compute ground-truth scores, as compared to using just evaluators within the evaluators' background(s). While Table 2.7 shows that the "experts" had higher average correlations, none of the correlation coefficients were significantly different (all $p > 0.1$), using a difference in correlation coefficients test that transformed the coefficients with the Fisher Z-transform. As a result, we considered all evaluators.

We chose three different learning problems, meant to show how well the system could do in three typical scenarios. In all scenarios, we trained and tested the system using leave-one-child-out cross-validation, i.e., trained the system on 41 children and tested it on the held-out child, and repeated this process for all 42 children. In the first scenario, we trained the system on an individual evaluator's scores and tested on the same evaluator's held-out score. Scenario 1 is a test for how well the system can predict a single evaluator's scores if trained on that evaluator. In scenario two, we predicted individual evaluator's scores using ground-truth scores to train the system. In this scenario, we computed a ground-truth score for each child by taking the mean score across the 10 held-out evaluators. Scenario 2 is a test for how well the system can predict single evaluator's scores if trained on a bank of held-out evaluators; scenario 2 is analogous to testing how much an evaluator agrees with "off-the-shelf" assessment tools trained on a group of different evaluators. In the third scenario (and the only one we did in our previous work [23, 24]), we predicted ground-truth scores using these ground-truth scores to train the system. Therefore, scenario 3 is a test for how well the system can predict a bank of evaluators if that same bank of evaluators trains the system.

To validate our results, we chose three metrics. Pearson's correlation coefficient (2.10) is the primary metric. Equation (2.11) is the mean absolute error between vectors of scores, y_1 and y_2. Equation (2.12) is the maximum absolute error between the two vectors of scores, y_1 and y_2.

$$E_{mean}(y_1, y_2) \equiv \frac{1}{42} \sum_{i=1}^{42} \left| y_1^i - y_2^i \right| \tag{2.11}$$

$$E_{max}(y_1, y_2) \equiv max \left(\left| y_1^1 - y_2^1 \right|, \cdots, \left| y_1^{42} - y_2^{42} \right| \right) \tag{2.12}$$

Before running experiments, we calculated human agreement statistics for all three metrics. Table 2.8 shows the human agreement statistics between the 11 evaluators, calculated in two ways: 1) using pairwise comparisons between individual evaluators and 2) comparing individual evaluators to the ground-truth scores of the other 10 evaluators. The pairwise comparisons had lower agreement than the ground-truth comparisons for all three metrics (lower correlation, higher mean absolute error, and higher maximum absolute error).

For all three scenarios, we chose to use linear regression techniques because of their simplicity and interpretability. The choice of function estimation methods made particular sense for scenarios 2 and 3, where the trained dependent variable was quasi-continuous. We also chose to use regression techniques for scenario 1, even though the dependent variable is ordinal, in order to ensure the results across the three scenarios are comparable. We did not z-normalize the dependent variable in any of the three scenarios since it had no impact on performance and since knowledge of the mean and standard deviation of the evaluator's scores in a real-life scenario is not always practical to attain.

For all experiments, we used leave-one-child-out cross-validation to separate train and test sets. Optimal learning parameters and feature subsets (when applicable) were computed on each cross-validation train set separately by using leave-one-child-out cross-validation; we chose the parameter settings (feature subsets) that maximized correlation between the automatic predictions and the evaluators' scores. This cross-validation approach effectively made use of all labeled data and simultaneously ensured that we were testing the true predictive power of our features/methods.

We developed two baseline systems, based on token-level pronunciation assessment research, where pronunciation correctness is often solely considered. Both baselines use simple linear regression with single features. The first uses the mean of feature VER1, and the second uses the mean of feature VER8 (Table 2.6). These two features represent the fraction of words mispronounced by the child, as determined by the LEX and GOP pronunciation verification methods, respectively (Section 2.3.5). Therefore, the baseline methods test whether one-dimensional token-level assessments can be extended to high-level assessments by simply computing an average over the token-level assessments.

A logical extension to these baseline systems would be to use multiple linear regression with the full set of 576 child-level features. Equation (2.13) shows this linear model, where $\overline{y}$ is the centered (mean subtracted) vector of human scores, X is the matrix of child-level features, w is the vector of coefficient weights, and ϵ is a zero mean Gaussian random variable. The objective function J in this case is (2.14), and (2.15) is the analytical solution which minimizes J.

$$\overline{y} = Xw + \epsilon \tag{2.13}$$

$$J = \|\overline{y} - Xw\|^2 \equiv (\overline{y} - Xw)^T(\overline{y} - Xw) \tag{2.14}$$

$$w = (X^T X)^{-1} X^T \overline{y} \tag{2.15}$$

Due to multicollinearity in the feature set, the solution to the inverse in (2.15) would be numerically unstable. We addressed this problem by trying various feature selection methods that model the dependent variable as a linear combination of a sparse set of independent variables. Choosing a subset of the features implicitly filters out redundant, irrelevant, and/or noisy features and makes the model easier to interpret. To show the relative merits of each feature, we ran simple linear regression (SLR) with each child-level feature individually.

We next tried three feature selection methods within the linear regression framework: a forward selection method, stepwise linear regression, and the "lasso" (least absolute shrinkage and selection operator) [166]. Forward selection iteratively adds features that optimize Pearson's correlation coefficient (2.10). Stepwise regression is less greedy in that it can remove entered features if their coefficient's p-values become too large. The lasso algorithm finds a solution to the least-squares error minimization when adding a λ-weighted L1 regularization term to the objective function, as shown in (2.16). This penalizes solutions with large weight coefficients (which often occurs when features are correlated) and promotes sparse models. Thus, many of the weight coefficients will be identically zero. We implemented the lasso using the least angle regression (LARS) algorithm, since there is no analytical solution to the lasso objective function [61, 66]. Note that we must standardize the features to ensure the regularization term is applied equally to all features. We accomplished this by centering the feature matrix X and dividing by the standard deviation of each feature; this normalization is denoted in (16) as $\widetilde{X}$.

$$J = \left\| \overline{y} - \widetilde{X} w \right\|^2 + \lambda \left\| w \right\| \tag{2.16}$$

2.6 Results and Discussion

Table 2.9 shows the performance for the two aforementioned baseline methods, the performance of the best SLR features for each of the three feature types, and the performance for the three feature selection methods. Table 2.10 provides coefficient statistics and lists which features were selected in at least 20% of the 42 cross-validations for the best performing feature selection method in each of the three train/test scenarios. We see from these results that scenario 1 (training and testing on individual scores) is the hardest, followed by scenario 2 (training on ground-truth scores and testing on a held-out evaluator), followed by scenario 3 (training and testing on ground-truth scores). We can explain the relative difficulty of the three scenarios using the following high-level description. Individual evaluators' scores can be viewed as "noisy," due to the subjective nature of the assessment task. Averaging the evaluators' scores can be seen as a method to "de-noise" individual evaluators' scores. We get the best results in scenario 1, where we train and test on ground-truth ("de-noised") scores and the worst results when we train and test on individual ("noisy") evaluators' scores.

In Table 2.10, we see that the baseline methods (that used the means of VER1 and VER8), did not use the best features, since the mean of VER10 proved to be a better predictor of the children's overall reading ability in all three learning scenarios. VER10 combines VER1 and VER8 into one trinary verification feature (Table 2.6). When limited to one feature, this single verification feature achieved the best results in terms of all three metrics and for all three scenarios, compared with using a single fluency or speaking rate feature (Table 2.9).

Within each scenario, the automatic methods that used multiple features outperformed the single feature methods (including the two baselines) for all three metrics. For scenario 1, we achieved the best results in terms of correlation (2.10) and mean absolute error (2.11) using the lasso regression as a pre-processing feature selection algorithm and then training the coefficient weights using multiple linear regression; we achieved the best results in terms of maximum absolute error (2.12) using the lasso method to select features and train the weights. For scenario 2, we achieved the best results for all three metrics using forward feature selection. For scenario 3, we got equally good results with both the forward selection and stepwise linear regression methods. Forward linear regression most likely achieved the best results for Scenarios 2 and 3 because the resulting feature set included only two features, so a greedy forward selection process was sufficient and outperformed more complicated feature selection methods. On the other hand, for Scenario 1, the lasso algorithm provided a more robust objective function for the more difficult learning problem, and the average number of features selected at each cross-validation was much higher at 5.6. Thus, in this case, the forward selection algorithm was unable to robustly select this higher number of features. The stepwise linear regression method can be viewed as the middle ground, which explains why its performance generally fell between that of the forward selection and the lasso. Table 2.10 also shows that for scenarios 2 and 3, the forward selection algorithm chose the top performing verification and fluency features for almost all of the cross-validation folds. However, for scenario 1, the lasso algorithm selected a variety of features, depending on the evaluator.

Scenario 3 was the only one in which we achieved a significantly higher correlation coefficient, compared to the best baseline system ($z = 2.78$, $p = .005$). Fig. 2.7 shows performance (in terms of correlation) of the different automatic feature selection methods for all three learning scenarios, compared to the human agreement statistics computed earlier. For the human agreement in this plot, we show the pairwise inter-evaluator correlations in scenario 1, and the ground-truth correlations in scenarios 2 and 3. We see from this plot that we were able to achieve a comparable level of human agreement for scenario 1 with the lasso and linear regression learning method. The mean automatic performance correlation of 0.828 was actually higher than the average pairwise human evaluator correlation of 0.827, although this difference was not significant ($z = 0.014$, $p = 0.989$). This means that the system trained on a particular evaluator will agree with that evaluator about as much as other evaluators will agree with that evaluator. In scenario 2, the automatic performance improved, benefiting from being trained on the perceptions of multiple evaluators, but its average performance was less than human agreement in this scenario, since the scores being predicted were from a held-out evaluator (resulting in a mismatched train/test condition). For scenario 2, the human evaluators' scores were correlated with ground-truth scores with 0.899 correlation, which was not significantly higher than automatic correlation of 0.869 ($z = 0.609$, $p = 0.542$). In scenario 3, the automatic performance is greater than average human agreement, although not significantly ($z = 1.44$, $p = 0.151$). In this scenario, the automatic system had the benefit of having multiple evaluators to train the system and also a matched test set composed of the same evaluators.

Fig. 2.8 shows the automatic predictions for the best automatic system in scenario 3. The automatic predictions were inside the mean human errors for 34 out of 42 (81%) of the children. We ran a final experiment by re-running scenario 3 using random subsets of evaluators (ranging from 2 to 10 evaluators). Fig. 2.9 shows these results when using the forward selection and lasso/linear regression methods. Again, for this plot, we also show agreement between the human evaluators (comparing individual evaluators to the ground-truth scores of the other selected evaluators). We chose 10 random subsets of evaluators for each value of the number of evaluators chosen. We see from this plot that human agreement and automatic performance both improve as a function of the number of evaluators. More importantly, we see that automatic performance is relatively high, even when using multiple evaluators with just two evaluators. This shows that the system benefits from the joint modeling of evaluators with as few as two evaluators.

2.7 Incorporating Evaluator Variability

One weakness to our proposed approach in Section 2.4 was that we did not take into account the fact that the variability in ratings across evaluators was not constant for all children; evaluators were in complete agreement for some children and disagreed more for other children. This variable level of evaluator uncertainty could potentially be incorporated during model training. In addition, we will show that this *heteroscedasticity* in evaluators' subjective judgments (having a non-constant variance) violates an assumption of the least squares linear regression techniques proposed in [16]. We addressed this weakness in this section by employing generalized least squares linear regression methods that account for this "variable variability" in evaluators' scores across children [20].

Figure 2.10 is a plot of the mean and standard deviation in the overall reading ability scores assigned to each child, computed across all evaluators. We see from this figure that the mean scores ranged from 1.55 to 7, and the standard deviations ranged from 0 (all evaluators agreed for 2 of the 42 children) to 1.29. The lowest standard deviations occurred for the children with higher mean scores. This makes numerical sense for the children with mean scores greater than 6.5 because all evaluators assigned scores of 6 or 7. However, it can also be argued that these children are objectively *easier* to grade, since they spoke most of the words correctly and had few disfluencies.

On the other hand, evaluators tended to agree less for the children with more pronunciation errors and more disfluencies; these cues may have impacted the evaluators to differing degrees. Thus, it can be argued that it is more *subjective* to grade the children with the higher standard deviations. In particular, the child with the highest standard deviation (who was assigned scores that ranged from 2 to 7) pronounced almost all of the words correctly but sounded out each word beforehand; it is possible that some evaluators largely ignored these sound-out disfluencies, while others felt it was strong evidence that the child was not (yet) the most skilled reader.

While Figure 2.10 provides visual evidence that the evaluators' level of agreement varied across children, we also employed two statistical hypothesis tests for heteroscedasticity: Levene's test [116] and the Brown-Forsythe test [30]. For both tests, we could reject the null hypothesis of homoscedasticity in the evaluators' scores at the 5% significance level (Levene's: $p < 0.001$, Brown-Forsythe: $p < 0.05$). This validates our decision in this work to pursue generalized least squares linear regression methods, which do not assume the evaluators' overall scores have equal variance for each child.

Our goal in this section is to predict the overall reading ability scores from the *mean* evaluator (Figure 2.10) using the same features discussed in Section 2.3. We explain the baseline system in Section 2.7.1 and our proposed methods in Section 2.7.2. For all methods, we used leave-one-out cross-validation to separate training data (41 children) from the test child. We optimized all regression parameters (e.g., selected features, smoothing/tuning parameters) using another stage of leave-one-out cross-validation on each train set separately.

2.7.1 Least squares linear regression

The baseline learning method, least squares (LS) linear regression, was based on our previous work [16]. The problem is defined as:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.17}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector comprised of the mean evaluator scores for each child, X is the noiseless $n \times m$ feature matrix (with a $n \times 1$ ones vector appended to account for the intercept/offset term), $\boldsymbol{\beta}$ is the $m \times 1$ linear weight vector, and the $n \times 1$ residual vector

$\boldsymbol{\epsilon}$ is *assumed* to be homoscedastic. The optimal linear weights $\hat{\boldsymbol{\beta}}$ that minimize the sum of the squared residual, $\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2$, are:

$$\hat{\boldsymbol{\beta}}_{ls} = \left(X^{\mathsf{T}}X\right)^{-1}X^{\mathsf{T}}\boldsymbol{y} \tag{2.18}$$

Due to dimensionality issues and multicollinearity effects, we did not use all 576 features in X. Instead, we used sequential forward feature selection to iteratively select features and construct X that maximized Pearson's correlation between $\boldsymbol{y}$ and $X\hat{\boldsymbol{\beta}}$ on the train set. Two or three features were selected, depending on the cross-validation fold ($m = \{2, 3\}$). Therefore, $n > m$, and we never had the problem of an under-determined system.

2.7.2 Generalized least squares linear regression

The least squares solution shown in Equation 2.18 is only optimal when the assumption of homoscedasticity in $\boldsymbol{\epsilon}$ holds. However, since we showed that $\boldsymbol{y}$ is heteroscedastic, we see in Equation 2.17 that $\boldsymbol{\epsilon}$ too will be heteroscedastic. This led us to employing generalized least squares linear regression methods [39]. In this formulation, the optimal linear weights, in the least squares sense, are:

$$\hat{\boldsymbol{\beta}} = \left(X^{\mathsf{T}}\Omega X\right)^{-1}X^{\mathsf{T}}\Omega\boldsymbol{y}, \tag{2.19}$$

where Ω is a diagonal matrix, with diagonal elements $\Omega_{jj} = 1/\sigma_j^2$, where σ_j is the "true" standard deviation in the overall reading ability of child j; see [39] for a derivation. In this section, we estimated Ω in two ways: 1) by using the scores provided by the 11 evaluators, and 2) by iteratively estimating Ω from the prediction residuals. We refer to the former method as weighted least squares (WLS) and the latter method as feasible generalized least squares (FGLS)[2].

Equation 2.20 shows how we computed the WLS estimate of Ω, where $\tilde{\sigma}_j$ is the estimated standard deviation in the overall reading ability of child j, computed from the evaluators' scores (Figure 2.10), and C_w is a positive smoothing parameter:

$$\Omega_{wls} = \text{diag}\left(\frac{1}{\tilde{\sigma}_1^2 + C_w}, \cdots, \frac{1}{\tilde{\sigma}_n^2 + C_w}\right), \quad C_w > 0 \tag{2.20}$$

The WLS method has the benefit of requiring only one additional parameter, C_w, which is needed to avoid numerical problems for the case when all evaluators agree ($\tilde{\sigma}_j = 0$). C_w can also be viewed as a tuning parameter; as C_w is increased, the solution to $\hat{\boldsymbol{\beta}}_{wls}$ (Equation 2.19) tends to $\hat{\boldsymbol{\beta}}_{ls}$ (Equation 2.18). Figure 2.11 demonstrates the effectiveness of the WLS method in predicting the mean evaluator's overall reading ability scores for a large range of C_w values.

For the FGLS method, we iteratively estimated Ω. See Algorithm 1 for pseudocode of our implementation, which was based on [39]. We first found the WLS solution on the training data and computed the prediction residual vector, which were used to initialize the FGLS iteration process. At each FGLS iteration i, the residual vector was used to construct a new FGLS diagonal matrix Ω_i (step 8). The form of Ω_i is very similar to Ω_{wls} (Equation 2.20), except Ω_i is determined analytically from the trained model, while Ω_{wls} is computed from the evaluators' scores. The FGLS smoothing parameter, C_f, in step 8 of

[2] FGLS is also commonly known as iteratively reweighted least squares.

Algorithm 1 Feasible generalized least squares (FGLS)

```
Require: Training data (feature matrix: $X$, dependent variable: $\boldsymbol{y}$)
 1: Compute weighted least square (WLS) solution: $\hat{\boldsymbol{\beta}}_{wls}$
 2: Compute WLS residual column vector: $\boldsymbol{\epsilon}_{wls} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{wls}$
 3: Compute sum of squared residual: $E_{wls} = \boldsymbol{\epsilon}_{wls}^{\mathsf{T}}\boldsymbol{\epsilon}_{wls}$
 4: Initialize FGLS: $\hat{\boldsymbol{\beta}}_0 \leftarrow \hat{\boldsymbol{\beta}}_{wls}$, $\boldsymbol{\epsilon}_0 \leftarrow \boldsymbol{\epsilon}_{wls}$, $E_0 \leftarrow E_{wls}$
 5: Initialize FGLS iteration counter: $i \leftarrow 0$
 6: repeat
 7:     Increment FGLS iteration counter: $i \leftarrow i + 1$
 8:     Compute diagonal FGLS matrix:
        $\Omega_i = \text{diag}\left(\frac{1}{\epsilon_{i-1,1}^2 + C_f}, \cdots, \frac{1}{\epsilon_{i-1,n}^2 + C_f}\right), \quad C_f > 0$
 9:     Compute FGLS coefficients: $\hat{\boldsymbol{\beta}}_i = (X^{\mathsf{T}}\Omega_i X)^{-1} X^{\mathsf{T}}\Omega_i \boldsymbol{y}$
10:     Compute FGLS residual column vector: $\boldsymbol{\epsilon}_i = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}_i$
11:     Compute FGLS sum of squared residual: $E_i = \boldsymbol{\epsilon}_i^{\mathsf{T}}\boldsymbol{\epsilon}_i$
12: until $E_i \geq E_{i-1}$
```

the algorithm is analogous to the C_w term in Equation 2.20. We selected C_f using a grid search, choosing the value that maximized the Pearson's correlation between the diagonal entries of Ω_i and Ω_{wls}; this tuning method was used to avoid over-training and numerical issues. In steps 9 and 10 of Algorithm 1, new estimates for the FGLS linear weights $\hat{\boldsymbol{\beta}}_i$ were computed and a new residual vector was calculated. This iterative process was repeated until the sum of the squared residuals no longer decreased on the training data. After convergence, the trained model was then applied to the test data. We found the FGLS algorithm converged in 3 to 12 iterations, depending on the cross-validation fold.

For illustrative purposes, Figure 2.12 shows the performance of the FGLS method in predicting the mean evaluator's overall reading ability scores, as a function of the FGLS iteration. While we only attained a small gain in performance over WLS with respect to Pearson's correlation, we do get a larger relative boost in performance for the two secondary metrics used in [16] and listed in Section 2.4: the *mean* absolute error in predictions and the *maximum* absolute error in predictions (out of the 42 children). This suggests that the FGLS method helps improve the robustness in estimating the linear weight coefficients $\hat{\boldsymbol{\beta}}$ by starting from the WLS solution and iteratively incorporating uncertainty in the trained model.

2.7.3 Results and Discussion

Comparable results for the three learning methods, attained by selecting features and optimizing all learning parameters using cross-validation, are shown in Table 2.11. We see that both proposed methods (WLS and FGLS) equaled or outperformed the baseline LS method for all three performance metrics. While there were no significant differences in the correlation coefficients of the three methods, the incremental improvements achieved with the WLS and FGLS methods made their correlations significantly higher than the mean inter-evaluator agreement of 0.899 (Table 2.7), with both $p < 0.05$.

The WLS method, which directly modeled evaluators' variability across children, achieved a Pearson's correlation coefficient of 0.951 between the predicted scores and the mean evaluator's scores, a relative improvement of 0.53% over baseline LS linear regression. The best overall system for all three performance metrics was FGLS linear regression, with relative improvements over baseline LS linear regression of 0.63%, 2.5%, and 1.4% for the correlation, average absolute error, and maximum absolute error performance metrics, respectively. The FGLS method has the benefit of being initialized with the WLS solution and making further changes based on the heteroscedasticity of the residual from the trained model.

2.8 Conclusions

This chapter addresses the need for automatic literacy assessments by predicting high-level ratings of children's overall reading ability, based on their performance reading a list of words aloud. We chose to use a modeling scheme that linearly combined a sparse set of features that spanned the ones actual human evaluators said they used (pronunciation correctness, fluency, and speaking rate). The resulting multi-dimensional models implicitly weight the importance of the selected features and offer a more interpretive assessment than the more common token-level assessments. As part of this work, we developed methods to automatically detect mispronunciations and disfluencies on a development training set, using grammar-based automatic speech recognition.

The automatic models performed best when trained on a bank of evaluators and when the train and test set were matched. We showed we could improve the predictive power by incorporating variability in evaluators' uncertainty across children using generalized least squares linear regression. We hope the techniques proposed in this thesis can be applied to other learning problems that involve modeling the perceptions of multiple evaluators.

One area of future work is to take into account evaluator *reliability*, as opposed to treating each evaluator equally; this has been shown to be advantageous in the context of emotion classification [6]. The inter-evaluator agreement statistics listed in Table 2.7 vary for the 11 evaluators, so it is possible that some evaluators are more reliable than others. We may be able to predict the evaluators' scores better if we weighted the scores of the more reliable evaluators higher. Unfortunately, initial experiments that used evaluator reliability-weighted linear combinations of the scores (using the agreement statistics in

Table 2.7 as a measure of reliability) did not increase automatic prediction performance. Future research will experiment with other reliability metrics to find more robust ways of combining multiple evaluators' perspectives (e.g., by using data-dependent evaluator modeling as in [5]).

This type of automatic processing could be especially useful in a classroom environment, where the teacher or a number of teachers could train the system to mimic their grading trends. High-level assessments could then be used by teachers to ensure the children are learning at an appropriate rate and to help inform their lessons. This type of collaboration between technology and teachers could transform the classroom.

In the future, we would like to incorporate both audio and video information for a more realistic scoring scenario. We would also like to extend this high-level literacy assessment to other reading tasks. We imagine applying it within a framework that examines children's skills across various reading tasks, so as to provide teachers with analysis on areas in which a child might be excelling versus an area in which he/she may need more practice or instruction.

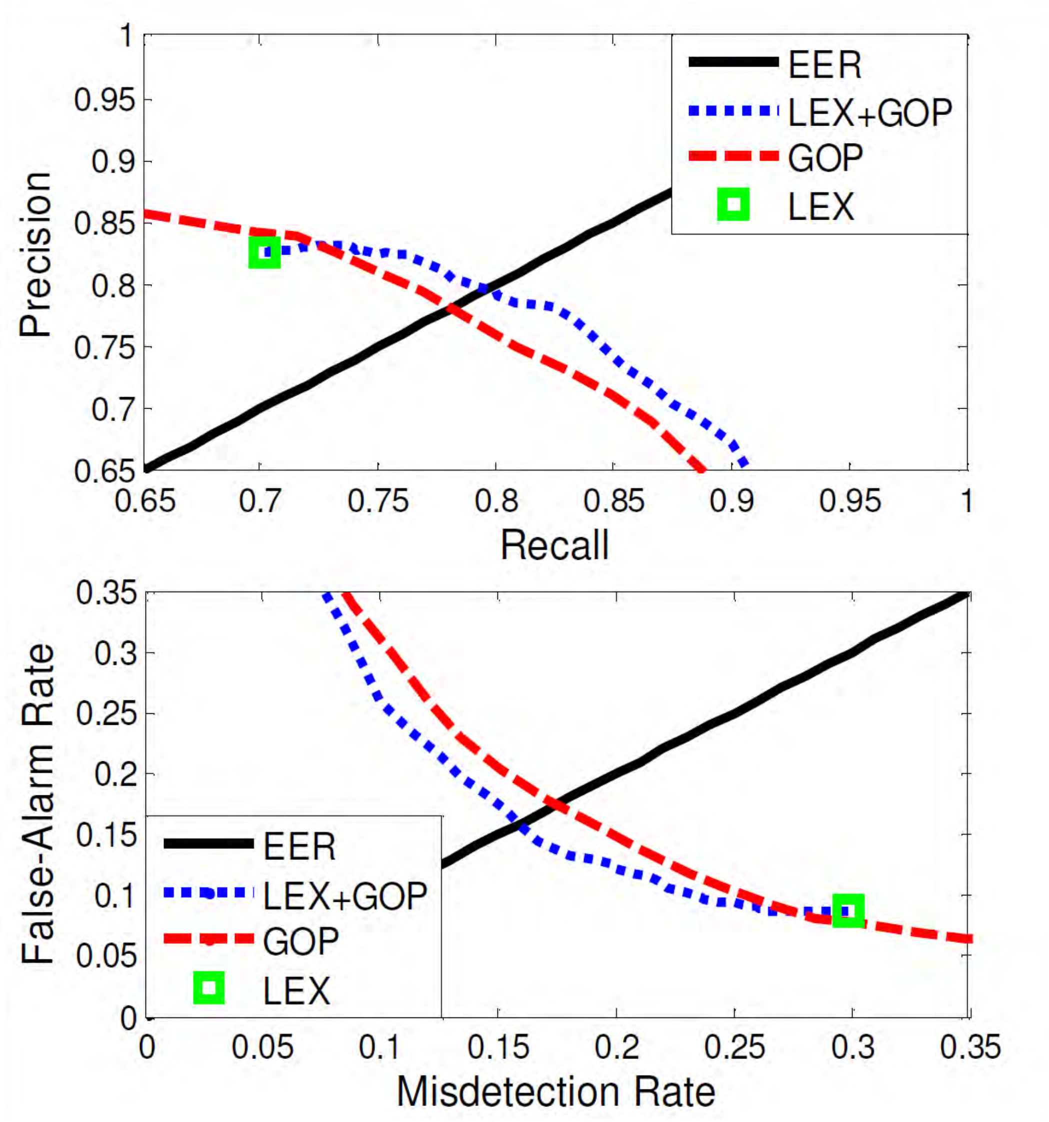

Figure 2.3: Performance of the three pronunciation verification methods (LEX, GOP, LEX+GOP). The GOP method performances are shown as the GOP score threshold is varied from -10 to 0. EER is the equal error rate for the displayed metrics.

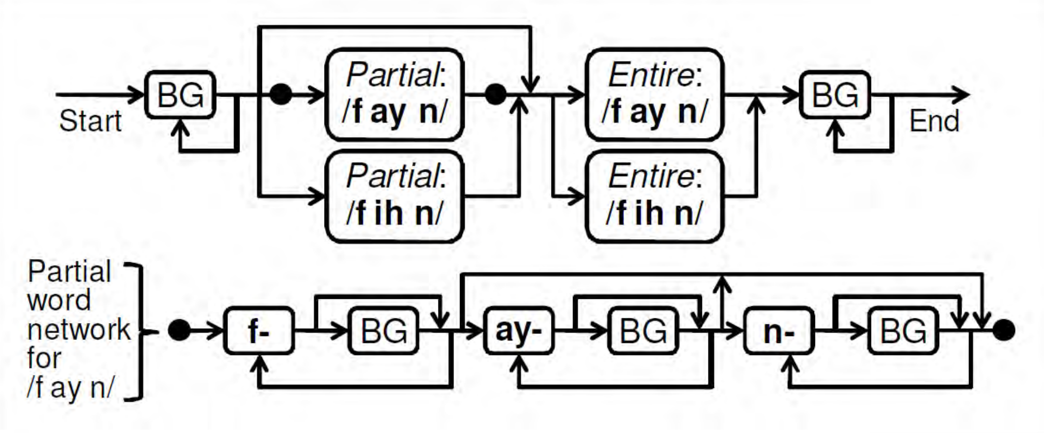

Figure 2.4: The stage 1 disfluency detection finite-state grammar (FSG) for the sample word, "fine," which has two entries in the dictionary (/f ay n/, /f ih n/). The FSG allows partial word manifestations of the target word to be recognized before a required forced-alignment of the entire target word. (BG is the background acoustic model.)

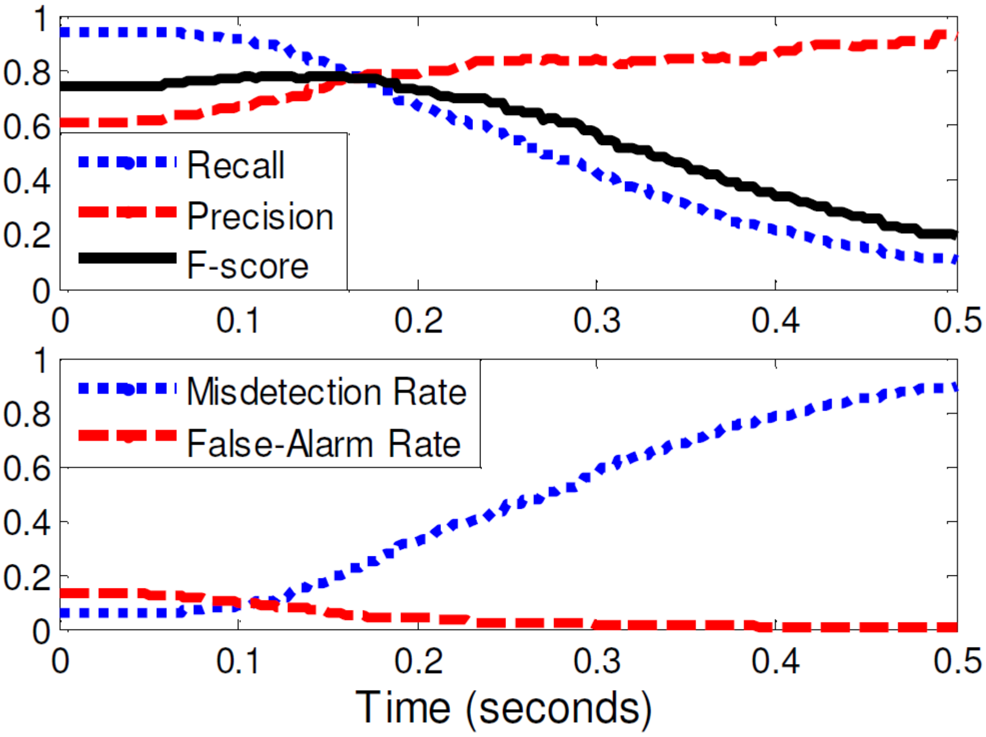

Figure 2.5: The performance of the stage 2 finite-state grammar (FSG) method as a function of the partial word length threshold (below which all partial words were rejected).

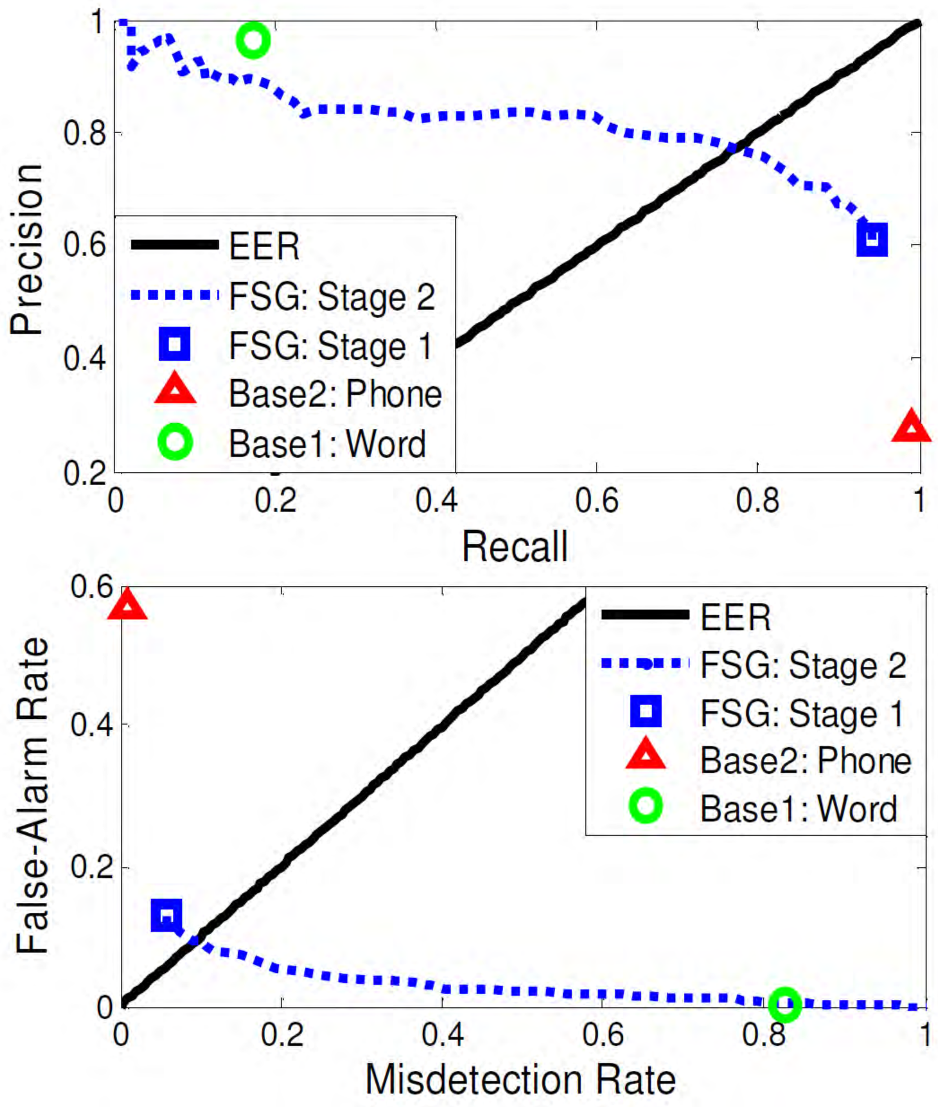

Figure 2.6: Performance of the two baseline systems (Base1 and Base2) and target word-specific finite-state grammar (FSG) procedure (stages 1 and 2). The FSG stage 2 performance is shown as the minimum partial word length threshold is varied from 0 to 2 seconds. EER is the equal error rate for the displayed metrics.

Name	Description	Domain
VER1	Was unacceptable pronunciation recognized?	$\{0, 1\}$
VER2	Was common reading error recognized?	$\{0, 1\}$
VER3	Was Spanish-related error recognized?	$\{0, 1\}$
VER4	Was garbage (GG) recognized?	$\{0, 1\}$
VER5	Was background/silence (BG) recognized?	$\{0, 1\}$
VER6	Log-likelihood of acceptable pronunciation	$(-\infty, 0]$
VER7	GOP(w) - see (2.7)	$(-\infty, 0]$
VER8	Reject(w) - see (2.8)	$\{0, 1\}$
VER9	2-stage verification method - see Section 2.3.3	$\{0, 1\}$
VER10	VER1 + VER8	$\{0, 1, 2\}$
FL1	Number of recognized partial words	$\{0, 1, \ldots\}$
FL2	Was at least one partial word recognized?	$\{0, 1\}$
FL3	Length of recognized partial words [s]	$[0, 5)$
FL4	Length of silence between partial words [s]	$[0, 5)$
FL5	Length of all silence recognized [s]	$[0, 5)$
FL6	FL3 + FL4	$[0, 5)$
FL7	FL3 + FL5	$[0, 5)$
FL8:FL12	Square root of FL3 through FL7	$[0, \sqrt{5})$
SR1	Utterance length [s]	$(0, 5]$
SR2	Target word start time [s]	$[0, 5)$
SR3	Target word end time [s]	$(0, 5]$
SR4	Number of syllables spoken / (SR3 - SR2)	$(0, \infty)$
SR5	(SR3 - SR2) / Number of syllables spoken	$(0, 5]$
SR6	Number of phones spoken / (SR3 - SR2)	$(0, \infty)$
SR7	(SR3 - SR2) / Number of phones spoken	$(0, 5]$
SR8	Speech start time (partial word or target word)	$[0, 5)$
SR9	Speech end time	$(0, 5]$
SR10	Number of syllables spoken / (SR9 - SR8)	$(0, \infty)$
SR11	(SR9 - SR8) / Number of syllables spoken	$(0, 5]$
SR12	Number of phones spoken / (SR9 - SR8)	$(0, \infty)$
SR13	(SR9 - SR8) / Number of phones spoken	$(0, 5]$
SR14:SR26	Square root of SR1 through SR13	–

Table 2.6: Features extracted for each word in the test data (VER = verification, FL = fluency, SR = speaking rate). The temporal features have an upper bound of 5 seconds since this was the maximum time allotted per word. All GOP scores in this study were finite, since all phone probabilities were non-zero.

Evaluator	Pairwise Evaluator Correlation										Avg. Correlations			
(Background)	1	2	3	4	5	6	7	8	9	10	mean		ground-truth	
											intra	all	intra	all
1 (Naïve)											0.776	0.770	0.810	0.833
2 (Naïve)	0.70										0.767	0.803	0.808	0.874
3 (Naïve)	0.85	0.83									0.843	0.860	0.909	0.940
4 (Non-native)	0.72	0.70	0.84								0.813	0.780	0.850	0.844
5 (Non-native)	0.76	0.85	0.86	0.84							0.857	0.848	0.913	0.928
6 (Non-native)	0.82	0.84	0.89	0.86	0.91						0.880	0.868	0.944	0.949
7 (Non-nat., Ling.)	0.82	0.79	0.88	0.74	0.82	0.87					0.810	0.816	0.866	0.886
8 (Linguist)	0.69	0.86	0.84	0.73	0.87	0.86	0.73				0.777	0.814	0.801	0.888
9 (Linguist, Expert)	0.79	0.82	0.88	0.76	0.83	0.83	0.88	0.83			0.860	0.840	0.923	0.916
10 (Expert)	0.77	0.80	0.86	0.79	0.86	0.86	0.81	0.87	0.86		0.857	0.837	0.886	0.913
11 (Expert)	0.78	0.84	0.87	0.82	0.88	0.88	0.82	0.86	0.87	0.86	0.863	0.844	0.895	0.922
										Avg:	0.828	0.827	0.873	0.899

Table 2.7: Pairwise evaluator correlations between the 11 evaluators (Naïve = native English speakers with no background in linguistics or children's literacy, Non-Native = non-native English speakers with an engineering background in speech-related research, Linguist = taken at least two graduate-level linguistics courses, Experts = more than a year working on children's literacy research). Average correlations were computed two different ways ("mean" and "ground-truth") and across two different groupings of evaluators ("intra" and "all"). "Mean" is the average pairwise evaluator correlation, and "ground-truth" is the correlation between an evaluator's scores and the averaged scores of the other evaluators. "Intra" calculations compare evaluators with the same background(s), while "all" calculations compare all evaluators' scores.

Evaluator Domain	Mean (Standard Deviation)		
	Corr	Emean	Emax
Pairwise	0.827 (0.032)	0.810 (0.180)	2.800 (0.701)
Ground-Truth	0.899 (0.038)	0.624 (0.137)	2.227 (0.388)

Table 2.8: Human agreement statistics for the 3 metrics (2.10)-(2.12).

Scenario:Method	Mean (Standard Deviation when applicable)		
	Corr	Emean	Emax
1:Base1 (VER1)	0.734 (0.062)	0.914 (0.106)	2.880 (0.358)
1:Base2 (VER8)	0.746 (0.048)	0.930 (0.121)	2.682 (0.475)
1:SLR (best VER)	0.769 (0.065)	0.882 (0.072)	2.610 (0.632)
1:SLR (best FL)	0.748 (0.054)	0.895 (0.139)	3.041 (0.480)
1:SLR (best SR)	0.705 (0.105)	0.924 (0.197)	3.385 (0.799)
1:Forward LR	0.792 (0.074)	0.815 (0.160)	2.659 (0.700)
1:Stepwise LR	0.805 (0.055)	0.786 (0.143)	2.852 (0.722)
1:Lasso	0.807 (0.087)	0.814 (0.223)	2.467 (0.565)
1:Lasso, then LR	0.828 (0.070)	0.721 (0.153)	2.549 (0.560)
2:Base1 (VER1)	0.741 (0.053)	0.968 (0.111)	3.044 (0.376)
2:Base2 (VER8)	0.756 (0.044)	0.970 (0.107)	2.763 (0.687)
2:SLR (best VER)	0.812 (0.041)	0.856 (0.084)	2.510 (0.643)
2:SLR (best FL)	0.731 (0.051)	0.979 (0.137)	3.345 (0.505)
2:SLR (best SR)	0.724 (0.062)	0.975 (0.175)	3.374 (0.554)
2:Forward LR	0.869 (0.038)	0.712 (0.138)	2.407 (0.520)
2:Stepwise LR	0.861 (0.035)	0.730 (0.133)	2.589 (0.703)
2:Lasso	0.851 (0.041)	0.846 (0.139)	2.544 (0.552)
2:Lasso, then LR	0.854 (0.037)	0.753 (0.125)	2.526 (0.495)
3:Base1 (VER1)	0.809	0.735	2.405
3:Base2 (VER8)	0.822	0.743	1.909
3:SLR (best VER)	0.888	0.596	1.601
3:SLR (best FL)	0.799	0.759	2.762
3:SLR (best SR)	0.783	0.789	2.858
3:Forward LR	0.946	0.365	1.594
3:Stepwise LR	0.946	0.365	1.594
3:Lasso	0.925	0.535	1.837
3:Lasso, then LR	0.940	0.414	1.636

Table 2.9: Automatic performance for the three scenarios described in Section 2.4. The methods above the dotted line use single features, and the ones below use multiple features. The numbers in red are the best performance achieved for the three scenarios.

Scenario:Method	Feature	% Folds	Coefficient stats	
			M	SD
1:Base1 (VER1)	Mean(VER1)	–	-0.755	0.051
1:Base2 (VER8)	Mean(VER8)	–	-0.771	0.042
1:SLR (best VER)	Mean(VER10)	–	-0.851	0.012
1:SLR (best FL)	Uquart(FL12)	–	-0.801	0.036
1:SLR (best SR)	Uquart(SR14)	–	-0.771	0.055
1:Lasso, then LR	Range(VER7)	50.9	0.140	0.129
	Mean(VER7)	44.4	0.258	0.267
	Iquart(SR2)	38.1	-0.314	0.203
	Uquart(FL12)	31.2	-0.142	0.116
	Mean(VER6)	27.9	0.199	0.149
	Lquart(FL2)	21.3	-0.276	0.147
2:Base1 (VER1)	Mean(VER1)	–	-0.824	0.009
2:Base2 (VER8)	Mean(VER8)	–	-0.839	0.007
2:SLR (best VER)	Mean(VER10)	–	-0.898	0.005
2:SLR (best FL)	Uquart(FL12)	–	-0.852	0.006
2:SLR (best SR)	Uquart(SR14)	–	-0.829	0.009
2:Forward LR	Mean(VER10)	99.1	-0.604	0.017
	Uquart(FL12)	97.0	-0.442	0.019
3: Base1 (VER1)	Mean(VER1)	–	-0.825	0.007
3: Base2 (VER8)	Mean(VER8)	–	-0.840	0.006
3: SLR (best VER)	Mean(VER10)	–	0.899	0.004
3: SLR (best FL)	Uquart(FL12)	–	-0.852	0.005
3: SLR (best SR)	Uquart(SR14)	–	-0.829	0.006
3:Forward LR	Mean(VER10)	100.0	-0.605	0.012
	Uquart(FL12)	97.6	-0.442	0.013

Table 2.10: Statistics of the standardized coefficients for the baseline, single feature, and best performing feature selection methods.

System	*Performance metric*		
	Corr	$\lvert E\rvert_{avg}$	$\lvert E\rvert_{max}$
Least Squares (LS) – Baseline	0.946	0.365	1.601
Weighted Least Squares (WLS)	0.951	0.364	1.601
Feasible Generalized LS (FGLS)	**0.952**	**0.356**	**1.579**

Table 2.11: Performance, in terms of the 3 metrics, of the 3 proposed systems: baseline least squares (LS), weighted least squares (WLS), and feasible generalized least squares (FGLS).

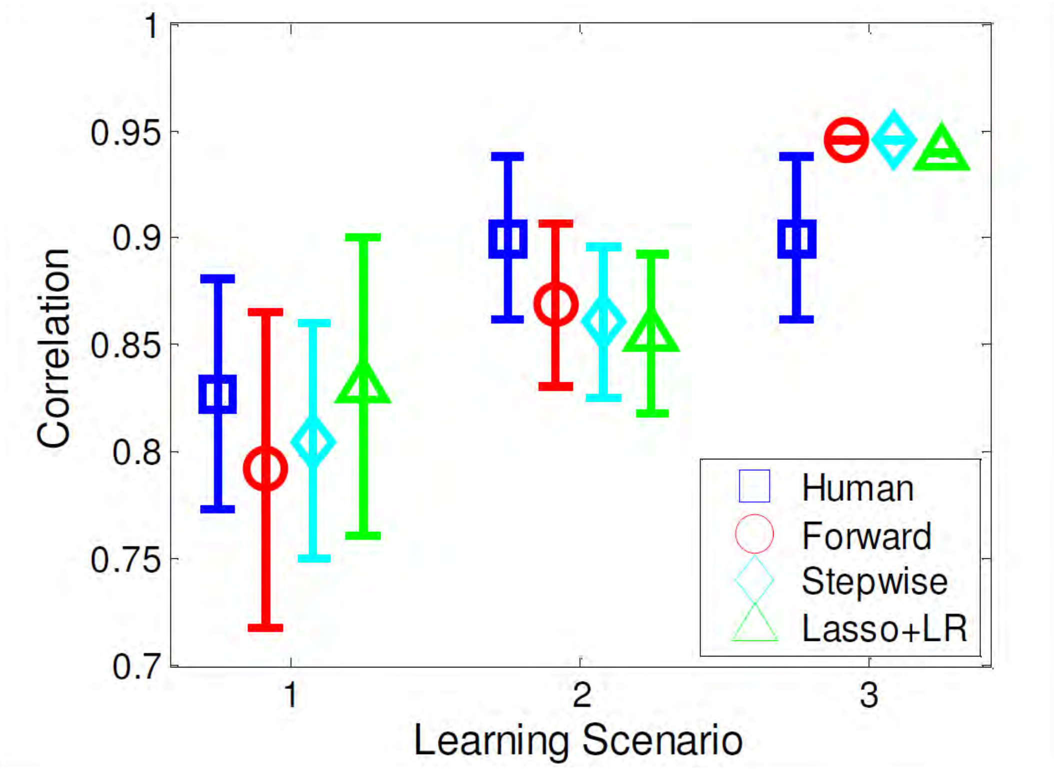

Figure 2.7: Mean and standard deviation of human evaluator agreement compared to the automatic performance for the three feature selection methods: forward selection, stepwise regression, and the lasso followed by linear regression.

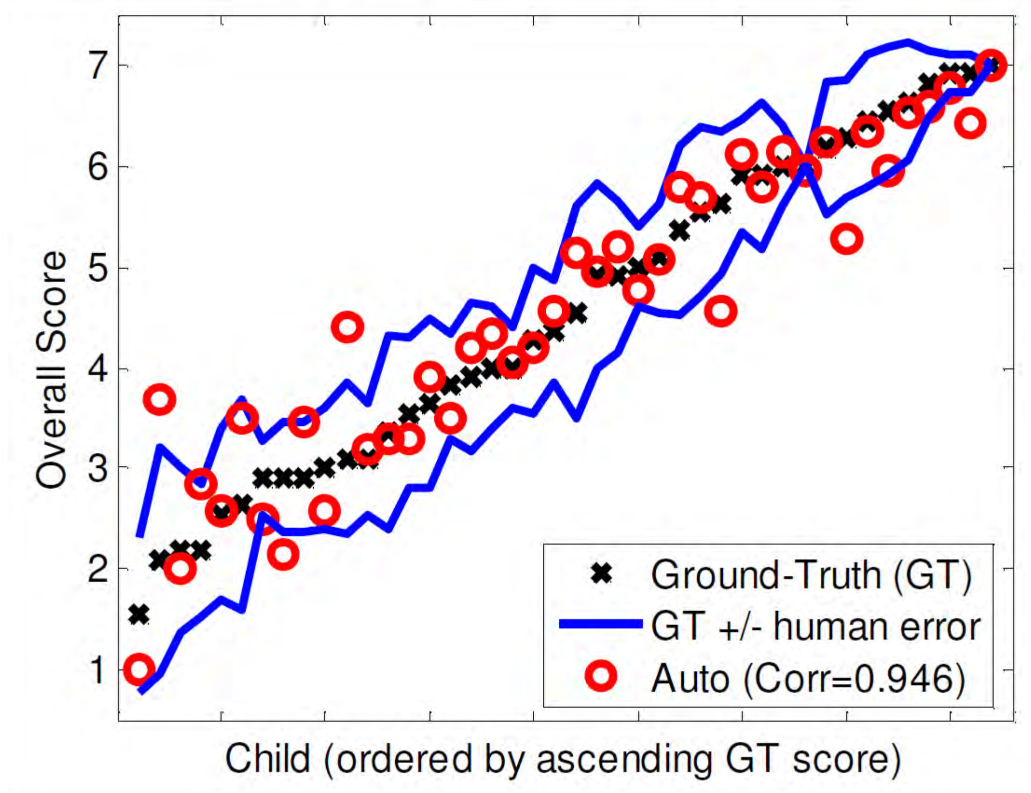

Figure 2.8: Linear regression results when using features selected using forward selection for scenario 3. "Human error" is the mean absolute difference from the ground-truth (GT) to held-out evaluators' scores.

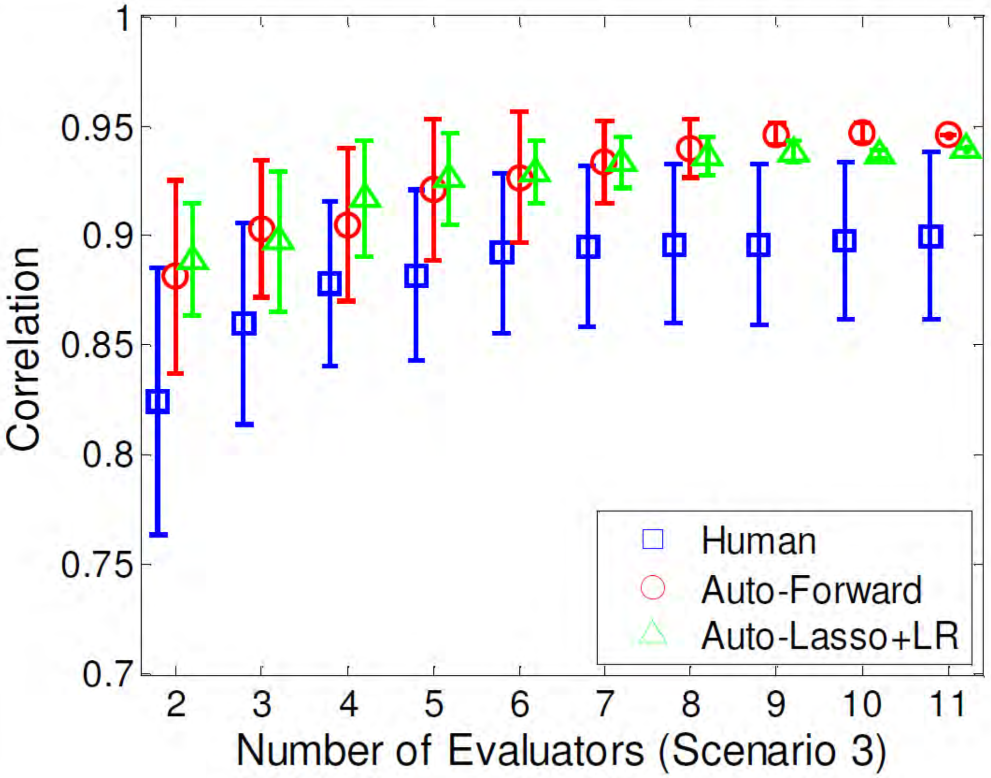

Figure 2.9: Correlation between predictions and evaluators' scores for learning scenario 3 as a function of the number of evaluators used to compute the ground-truth scores. It shows that both human agreement and automatic performance increase as the number of evaluators increases. Automatic performance with nine or more evaluators is significantly higher than with two evaluators ($z = 1.94$, $p = 0.048$).

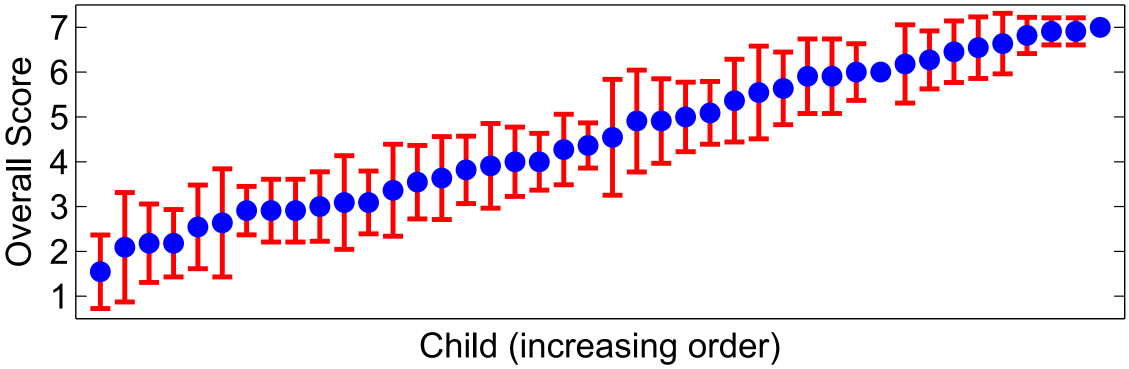

Figure 2.10: The mean and standard deviation in the overall scores assigned to each child, computed across all 11 evaluators.

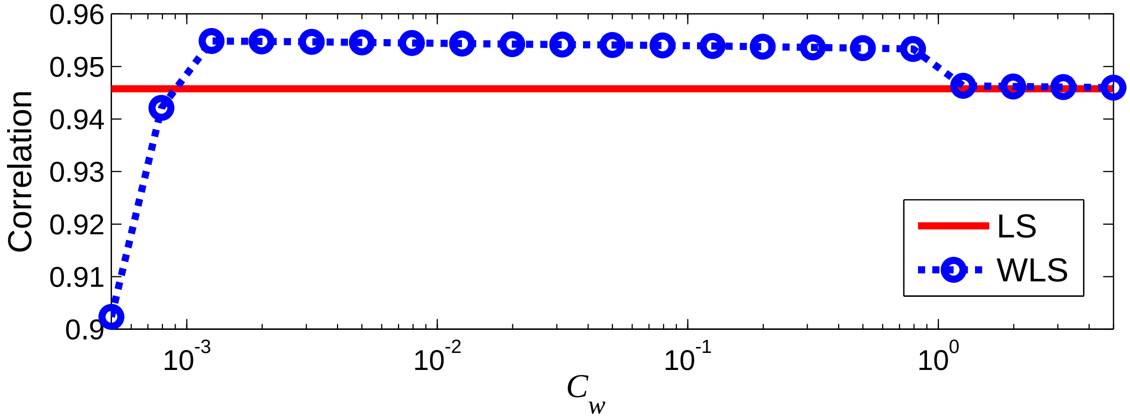

Figure 2.11: Performance, in terms of Pearson's correlation, in predicting the children's overall reading ability using the weighted least squares (WLS) method, as a function of the tuning parameter C_w.

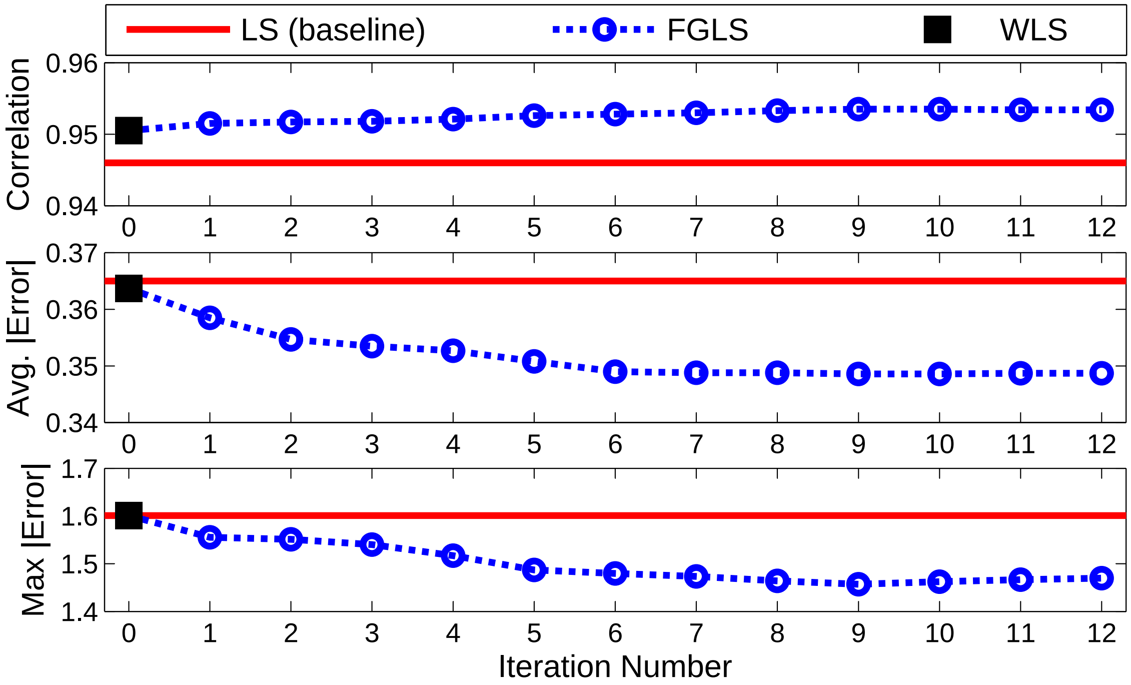

Figure 2.12: Performance, in terms of the 3 metrics, of the 3 proposed systems: baseline least squares (LS), weighted least squares (WLS), and 12 iterations of feasible generalized least squares (FGLS).

Chapter 3

Couples Therapy Research

3.1 Introduction

[1] In psychology and psychiatry, behavioral observation is essential for diagnosis for children and adults, and it is also a means for monitoring change during psychotherapy, where both therapist and client engage in, and respond to, continuous, albeit usually unsystematic, behavioral observation. The importance of observable behavior for researchers and therapists is borne of the fact that behavior is typically the best objective measure of psychologically relevant phenomena available. Self-reports of even obvious behaviors can be notoriously unreliable [140].

Although most observation in psychological and psychiatric practice has been unsystematic, systematic observational research has been central to numerous intra- and interpersonal psychological problem domains including depression [9], bi-polar disorder [75], anxiety [12], schizophrenia [31], autism [107], alcoholism [160], domestic aggression [119], and marital distress [96]. In each of these areas, observational research has identified

[1] This research was supported in part by the National Science Foundation and the Viterbi Research Innovation Fund. Special thanks to the Couple Therapy research staff for collecting, transcribing, coding, and sharing the data.

behaviors exhibited by individuals who suffer from such problems (and behaviors exhibited by family members and loved ones of afflicted individuals) that are associated with increased symptomatology and reoccurrence of disorders.

Behavioral observation has been used with considerable success in the study and treatment of intimate relationships. Current theory suggests, and recent empirical findings validate [81, 103], that spouses' behavior is a central and defining aspect of intimate relationships that links broad cultural factors, longstanding life experiences, and current stressors to the stability and quality of marital relationships.

However, the methods used in behavioral observation do present some challenges. To test research hypotheses, psychology and other fields in the behavioral sciences oftentimes rely heavily upon observational coding of audio-video data; for example, in family studies research, psychologists use a variety of established coding standards describing characterizations of specific behavior patterns of interest that guide human annotation of data [120]. This manual coding is a costly and challenging process. First, a detailed coding manual must be designed, which can be a complex iterative task [108].

After the creation of an appropriate coding manual, multiple coders, each with his/her own biases and limitations, must be trained in a consistent manner on held-out but representative data. In some cases, coders must meet a predetermined minimum level of agreement with a "gold-standard" coder on training data before they can code real data. To avoid coder drift, some coding protocols require coders to be evaluated periodically and retrained if necessary [108]. In addition, for longitudinal studies lasting several years, it is usually only feasible to have disjoint sets of coders, which adds another source of variability to the resulting coded data.

The actual coding process can be mentally straining and inefficient. Multiple coders oftentimes code the same data to allow for the computation of both code reliability and inter-rater reliability. Each coder observes the audio-video data and marks relevant behavioral phenomena according to the coding manual (e.g., in continuous time, in quantized time intervals, at the session-level). The complexity of the coding process determines the speed at which data can be coded, with more complex protocols taking orders of magnitude longer than real-time (e.g., [97]). To prevent evaluator fatigue, coders are often limited to coding for short periods of time in one sitting. Overall, the coding process is limited by the inherent subjective and qualitative nature of human descriptions on human behavior.

Technology has the potential to aid in coding human behavioral data. Computers are better suited to track and quantify certain behavioral phenomena that may be challenging, or even impossible, for humans to do. For example, whereas a human observer might have a qualitative idea of how a speaker's pitch may be changing, engineering algorithms can estimate and track the pitch of a speaker using quantitative methods at fine temporal granularities. Pitch, and other low-level descriptors (LLDs) of human behaviors [152], can be extracted using well-developed signal processing methods, which in turn can be mapped to relevant high-level human behavior via machine learning algorithms.

Computer technology has the advantage of automatically analyzing data in a consistent, repeatable manner. In addition, computational algorithms can be incrementally improved, benefiting from more data and improved methodologies. Another obvious advantage of computer technology is that it will not fatigue. Finally, whereas current human behavioral methods are not scalable to coding large amounts of data over long periods of time, computer technology is highly scalable. Technology can also be modularized, with separate algorithms specializing in modeling specific human behaviors, which could make the technology adaptable from one domain of research to another distinct but overlapping domain.

Our aim in this chapter is to augment the observational power of the researcher and therapist with novel computational tools and techniques. Specifically, we explore the power of objective signal-based measures (speech-derived audio cues), extracted during real marital discussions, in predicting perceptual observations made by evaluators trained on a manual human behavioral coding system. Thus, our goal is to emulate *human* evaluators observing *human* behavior.

This research is part of a growing field, behavioral signal processing (BSP), aimed at better connecting the behavioral sciences with signal processing methods. Traditional signal processing research (e.g., speech recognition, face/hand tracking) concentrated on modeling more objective human behaviors (e.g., "what was spoken?"). BSP builds upon traditional engineering tools and methods to model more abstract human behaviors in realistic scenarios that are especially relevant in psychology and related fields (e.g., the question "is one spouse blaming the other?" in a marital therapy session).

Significant work related to BSP has concentrated on extracting human-centered information from audio-video signals, including social cues [174], affect and emotions [86, 113, 153, 181], and intent [100]. The increased push to analyze realistic human interactions and naturalistic data (as opposed to acted or artificially constrained data) is most evident in the affective computing and emotion recognition communities [33, 38, 57, 58, 62, 63].

In this chapter, we apply the basic ideas of BSP using the Couple Therapy corpus [45], discussed in detail in Section 3.2 [17]. This corpus consists of recordings of a husband and wife spontaneously discussing a problem in their relationship. Each spouse's behavior was manually coded with a number of session-level codes (e.g., level of blame expressed, global positive affect). In [18], we showed that we could extract speech acoustic features that separated spouses' extreme behaviors significantly better than chance for three of the six behavioral codes we analyzed. In [109], we developed quantitative methods to model prosodic *entrainment* behavior between the spouses; couples rated as behaving more positive were found to have significantly higher levels of prosodic entrainment compared to couples rated as being more negative. In addition, the entrainment features were able to discriminate positively rated interactions from negatively rated ones.

This chapter represents an extension of [18], in which we analyze the same corpus. In this chapter, we improved upon our speaker segmentation method, which allowed us to analyze a larger percentage of the data in the corpus. We also took greater care in normalizing feature streams to combat variable acoustic conditions and speaker-dependencies. In addition, we experimented with new acoustic feature types and new techniques to map these features from the frame-level to the session-level. Finally, we compared various machine learning techniques to automatically predict the behavioral codes for the spouses. These extensions produced an absolute improvement of 3.95% in classifying the six behavioral codes, compared to the best results reported in [18].

Section 3.2 describes the Couple Therapy corpus, and Section 3.3 provides a methodological overview. We explain how we pre-processed the data in Section 3.4. Section 3.5

discusses the acoustic features we extracted to model the spouses' behavior, while Section 3.6 describes the learning methods and algorithms used to predict the spouses' behavioral codes. The results are presented and discussed in Section 3.7, with fusion experiments and results discussed in Section 3.8, and the conclusions and intended future work are provided in Section 3.9.

3.2 Couple Therapy corpus

The original study that produced the data we refer to as the Couple Therapy corpus was a multi-year, multi-university collaboration between researchers in the department of psychology at the University of California, Los Angeles and the University of Washington [45]. The main purpose was to test the efficacy of integrative behavioral couple therapy (IBCT) [47] versus traditional behavioral couple therapy (e.g., [11]) for treating severely and stably distressed couples who were not likely to benefit from other forms of couple therapy. This study became the largest longitudinal, randomized control trial of psychotherapy for severely and stably distressed couples and led to a number of psychology publications [10,43–45]. Based in large part on the success of IBCT as documented in these publications, IBCT is currently one of only four empirically supported interventions for relationship distress.

One hundred and thirty-four seriously and chronically distressed couples (all male-female pairs) were recruited in Los Angeles, California (71 couples) and Seattle, Washington (63 couples) and randomly split between the two couple therapy conditions. The recruitment inclusion criteria included: the couples being legally married and living together, both spouses speaking fluent English, being between the ages of 18 and 65, and having at least a high school education or its equivalent.

Recruited couples were married a mean of 10.0 years ($SD = 7.60$) at the beginning of the study. The mean age of the recruited wives was 41.6 years ($SD = 8.59$), and the mean age of the husbands was 43.5 years ($SD = 8.74$). The mean number of years of education was 17.0 for both the wives and husbands ($SD = 3.23$ for wives, $SD = 3.17$ for husbands). The majority of the participants were Caucasian (wives: 76.1%, husbands: 79.1%); other well-represented ethnicities included African American (wives: 8.2%, husbands: 6.7%), Asian or Pacific Islander (wives: 4.5%, husbands: 6.0%), and Latina/Latino (wives: 5.2%, husbands: 5.2%).

Each couple received up to 26 sessions of therapy over the course of one year. As part of the study, research staff had couples select two current, serious relationship problems, one chosen by each partner, and then had them engage in two dyadic discussions in which they were instructed to try to understand and resolve these respective relationship problems. There was no therapist or research staff present during these sessions, and the couple interacted for ten minutes about the wife's chosen topic and ten minutes about the husband's chosen topic; these two ten-minute sessions were considered separate and analyzed separately.

The problem-solving interactions were recorded at three points in time across the study: pre-therapy, the 26-week assessment, and the two-year post-therapy assessment. The audio-video data consist of a split-screen video (704x480 pixels, 29.97 fps) and a single channel of far-field audio recorded from the videocamera microphone (16 kHz,

Manual	Codes
SSIRS	global positive affect, global negative affect, use of humor, sadness, anger/frustration, belligerence/domineering, contempt/disgust, tension/anxiety, defensiveness, affection, satisfaction, solicits partner's suggestions, instrumental support offered, emotional support offered, submissive or dominant, topic a relationship issue, topic a personal issue, discussion about husband, discussion about wife
CIRS	acceptance of other, blame, responsibility for self, solicits partner's perspective, states external origins, discussion, clearly defines problem, offers solutions, negotiates, makes agreements, pressures for change, withdraws, avoidance

Table 3.1: A list of the 32 codes in the two human behavioral coding systems: Social Support Interaction Rating System (SSIRS) and Couples Interaction Rating System (CIRS).

16-bit). Since the data were originally only intended for manual coding, the recording conditions were not ideal for automatic analysis; the video angles, microphone placement, and background noise varied across couples and across sessions.

The audio-video recordings in the original study were used to manually code each spouse with relevant high-level behavioral information. Two separate rating systems ("coding manuals") were developed and used. Both were designed for use by naïve raters who were fluent in English and have a layperson's understanding of human interaction [157]. The Social Support Interaction Rating System (SSIRS) measured both the emotional content of the interaction as well as the topic of conversation [99]. It consisted of 19 questions ("codes") across four categories: affectivity, dominance/submission, features of the interaction, and topic definition. The Couples Interaction Rating System (CIRS) consisted of 13 codes and was specifically designed for coding problem-solving discussions [93]. All 32 codes had written guidelines and were on an integer scale from 1 ("none/not at all") to 9 ("a lot"). Table 3.1 lists the 32 codes in the two coding manuals.

Multiple coders rated each session (one set of 32 codes for *each* spouse) after watching the video at most two times. The number of coders per session ranged from 2 to 12, with 91.1% of the sessions being rated by 3 or 4 evaluators. Evaluator judgments were based on observation of the entire interaction and were at the session-level; no finer-grained codes were attained (e.g., utterance-level, turn-level). Evaluators were told to focus on one spouse (the "target spouse") when observing each interaction. They were encouraged to use information in both verbal and nonverbal channels when rating the spouse and to take into account both the frequency and intensity of particular behaviors, as well as the context in which they occur.

All coders were undergraduate students at the University of California, Los Angeles. They each underwent a training period to give them a sense for what was typical behavior and to help standardize the coding process. First, the coders rated acted videos of couples that exemplified low and high ratings of the codes. Then, coders compared their ratings with those of expert psychologists and discussed the differences. Evaluators began coding the real data once they demonstrated a reasonable level of reliability with the expert's ratings; inter-rater reliability varied depending on the code, as exemplified in Table 3.2 and explained in further detail in [158]. Typically the training process took approximately 15 hours. Evaluators continued to attend weekly two-hour training meetings to prevent drift and to ensure high reliability [157]. In total, 37 individual coders were trained across the two coding systems. It should be noted that disjoint sets of coders were used for the two coding manuals (a coder was only trained to rate the SSIRS or the CIRS), but coders rated couples across time periods.

As part of the original study, the sessions were manually transcribed for the purpose of analyzing the language use of each spouse [4, 10, 178]. They used the IBM ViaVoice speech transcription software, and the data took, on average, three to six times real-time to transcribe. The resulting word-level transcriptions were chronological, with the speaker explicitly labeled for each word (husband or wife). Nonverbal communication was marked in the transcriptions (e.g., laugh, sigh, throat clear, long pause). Spoken names and other proper nouns were de-identified in the transcriptions for privacy reasons, and transcribers also marked regions in which they could not understand the speech; in total, 0.98 percent of the words were either de-identified or unknown. In portions with overlapping speech, transcribers attempted to separate out words from each speaker, but regions of speech overlap were not explicitly marked. No timing information was provided in the transcriptions.

There are 574 ten-minute sessions with corresponding transcriptions in the Couple Therapy corpus. Five of these sessions were missing the codes from the two psychology rating systems. This left 569 coded sessions, totaling 95.8 hours of data across 117 unique couples.

3.3 Methodology overview

The Couple Therapy corpus provides a unique opportunity to test BSP methods and algorithms on data collected in an ecologically valid setting that meets the stringent standards used in behavioral science research. In addition, the size of the corpus makes it appealing for exploring data-driven BSP methods. Although the data quality is not optimal for automated processing, repeating the study to attain higher quality recordings of couples' interactions would entail a multi-year effort (for recruiting, subject scheduling, etc.). Furthermore, while the high variability in the recording conditions are a source of exaggerated noise, data quality variability is still present even in corpora collected with high-quality recording equipment, consistent sensor locations, and controlled acoustic/visual environmental conditions (e.g., [151]). We believe that analyzing this existing large corpus offers a veritable testbed for this domain of BSP research.

In this thesis, our goal was to provide analysis toward automatically learning a subset of the 32 codes using features derived from the audio signal. The following subsections explain the various design decisions we made. Section 3.3.1 describes the subset of codes we analyzed, and Section 3.3.2 explains the classification set-up for all experiments. Section 3.3.3 provides an overview of our methodology: data pre-processing, acoustic feature extraction, and supervised learning of the behavioral codes. Sections 3.4-3.6 provide more detailed descriptions of these three components, respectively.

3.3.1 Codes of interest

For clarity and to make the results comparable to our previous work [18], we chose to only analyze the following six codes with the highest inter-evaluator agreement: level of acceptance toward the other spouse (abbreviated "acc"), level of blame ("bla"), global positive affect ("pos"), global negative affect ("neg"), level of sadness ("sad"), and use of humor ("hum"). The Appendix provides the written guidelines for the six codes. It should be noted that each code measures how much that particular code occurred, *not*

Code	Code Correlation					Spouse Correlation	Agreement
	acc	*bla*	*pos*	*neg*	*sad*		
acc						0.647	0.751
bla	-0.80					0.470	0.788
pos	0.67	-0.54				0.667	0.740
neg	-0.77	0.72	-0.69			0.690	0.798
sad	-0.18	0.19	-0.18	0.36		0.315	0.722
hum	0.33	-0.20	0.47	-0.29	-0.15	0.787	0.755

Table 3.2: Correlation between each of the six codes, as well as the correlation between spouses' ratings and the inter-evaluator agreement for each of the codes. Pearson's correlation was the chosen metric.

how much the opposite of the code occurred. Therefore, it is possible for a spouse to receive high scores for both global positive affect and global negative affect.

Table 3.2 shows how the six codes are correlated, as well as the correlation between spouses' ratings and the inter-evaluator agreement for each of the six codes; Pearson's correlation coefficient was the chosen metric for all three computations. When computing the inter-code and spouse correlations, we used the mean evaluator scores for each instance. The agreement statistics were computed as the correlation between individual evaluator's scores and the mean scores of the other evaluators. All six selected codes had inter-evaluator agreement greater than 0.7; the remaining codes not analyzed in this dissertation had inter-evaluator agreement that ranged from 0.4 to 0.7.

We see in Table 3.2 that the *positive* codes (acc, pos, hum) were all positively correlated with each other, the *negative* codes (bla, neg, sad) were all positively correlated with each other, and the positive codes were negatively correlated with the negative codes; this agrees with intuition. We also see in Table 3.2 that the two spouses' behaviors were positively correlated for all six codes; this suggests that, on average, the interacting spouses displayed similar behaviors.

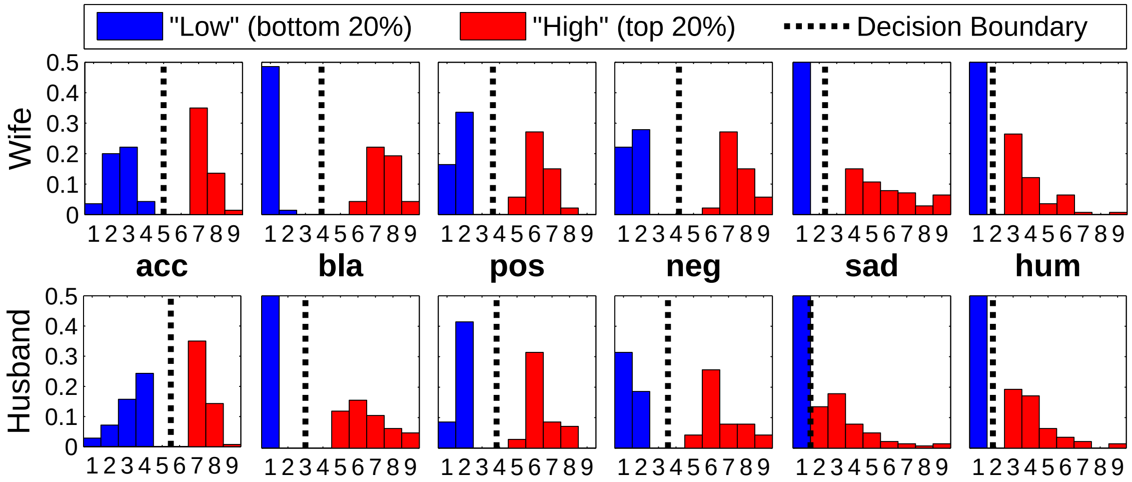

Figure 3.1: Normalized histograms of the extreme code scores for the wife (top) and husband (bottom). The "low" scores are in the bottom 20%, and the "high" scores are in the top 20%. The decision boundary was used to compute an upper-bound for automatic performance.

3.3.2 Classification task formulation

As described in Section 3.2, multiple coders rated each session (both spouses) for each behavioral code on a scale from 1 to 9. Thus, there are multiple ways to pose this learning problem for automatically predicting the behavioral code scores. Since there were disjoint sets of coders used, we ignored individual evaluator effects and treated each evaluator in the same manner.

Furthermore, we simplified the code learning problem by posing it as a binary classification problem, with equal-sized classes. We only analyzed sessions that had mean evaluator scores that fell in the top 20% ("high") and bottom 20% ("low") of the code range for both genders; see Figure 3.1. Therefore, our goal was to separate the *extreme* couples' behavior ratings for the six codes. A similar data-separating procedure was used in our previous paper [18] and in related work [100, 148]. This is a good starting point in trying to learn these subtle high-level behavioral codes.

Gender	*acc*	*bla*	*pos*	*neg*	*sad*	*hum*	AVG
Wife	96.7	99.6	98.5	98.6	93.9	96.5	97.3
Husband	96.7	98.1	97.4	98.0	84.9	97.1	95.4

Table 3.3: Upper-bound for automatic performance, computed as the percentage of individual coder scores that were within the decision boundary between the "low" and "high" code score groupings.

As shown in Figure 3.1, the "low" and "high" mean scores for the six codes are separable, i.e., the average coder scores for these extreme sessions do not overlap. However, this does *not* mean that individual coder scores were separable for this artificially created subset of the data. We produced an "upper-bound" for automatic performance by computing the level of individual human agreement with these low and high average score groupings. This was done by computing the percentage of individual evaluator scores (for the sessions in the top/bottom 20% of the code range) that fell within a code-specific decision boundary, which was placed halfway between the maximum "low" code score and the minimum "high" code score. These decision boundaries are shown in Figure 3.1, and the upper-bounds in code performance for the wife and husband are listed in Table 3.3. We see in this table that all of the upper-bounds were between 96% and 100%, except for level of sadness, which dipped as low as 84.9% for the husband; this is due to the fact that there is less separation between the extreme code scores (see Figure 3.1).

3.3.3 Classification system overview

See Figure 3.2 for a high-level system block diagram, which depicts the basic components of our methodology. First, we pre-processed the corpus by: 1) eliminating sessions that were too noisy, 2) automatically segmenting the sessions into single speaker regions, and 3) eliminating sessions for which we could not attain reliable speaker segmentation. These

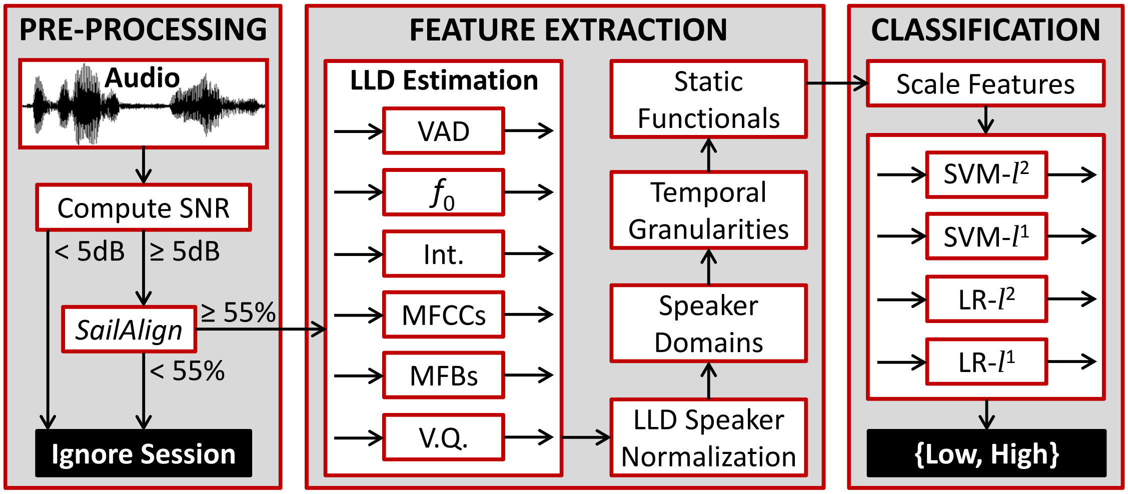

Figure 3.2: A system block diagram, illustrating the methodology taken in this thesis, from pre-processing the data and extracting acoustic features to classifying extreme instances of a particular code as low/high.

pre-processing steps were taken to eliminate sessions that were too noisy for the purpose of acoustic pattern recognition and to facilitate the extraction of spouse-specific acoustic features.

We estimated each session's average signal-to-noise ratio (SNR) and eliminated noisy sessions with an SNR less than 5 dB. To segment the corpus into single speaker regions, we used the available word-level transcriptions with speaker labels and *SailAlign* [104], software that implements a recursive speech-text alignment algorithm. To ensure we had at least a majority of the speech segmented for both spouses, we ignored all sessions for which we were unable to segment at least 55% of both the wife's and husband's words.

We extracted a set of low-level descriptors motivated by related work in both psychology and engineering, that along with their functionals resulted in a large set of over 40,000 features. This feature set was used to learn all six codes; code-specific features were not extracted. The features were *static functionals* (e.g., mean) of low-level descriptors (*LLDs*, e.g., intensity), computed over each *speaker domain* (e.g., wife regions) and at various *temporal granularities* (e.g., 0.5 s windows). Therefore, this feature extraction process mapped frame-level LLDs to session-level features that represented various acoustic properties of the spouses/interaction.

We extracted prosodic, spectral, and voice quality LLDs. The prosodic LLDs included: voice activity detector (VAD) estimates, speaking rate, fundamental frequency (f_0), and intensity. The spectrum-based LLDs included Mel-frequency cepstral coefficients (MFCCs) and log Mel-frequency bands (MFBs), and the voice quality (V.Q.) LLDs included jitter and shimmer. We normalized the raw LLD streams by speaker, since our goal was to train speaker-independent models for each of the behavioral codes.

We trained separate binary classifiers for each code. We experimented with two popular linear classifiers: support vector machines (SVM) with linear kernel and logistic regression (LR), and two types of regularization: l^2 and l^1. Regularization was applied to make the estimation of the feature linear weight coefficients more robust. In the case of l^1 regularization, a sparse solution is found, which facilitated an analysis on the relative importance of the features.

We used leave-one-*couple*-out cross-validation to separate training and test data; this was done to ensure that the reported results were representative of practical training conditions in which data from a couple would typically not be available. Note that we did not use leave-one-session-out cross-validation because some couples had more than one session in the top/bottom 20% for a particular code. All classifier parameters were optimized at each train/test fold using a second stage of 5-fold couple-disjoint cross-validation on the training data. To evaluate classifier performance, we pooled all the test class hypotheses and computed the percentage of correctly classified instances ("accuracy"). Chance baseline accuracy was 50%, since we have equal-sized classes.

We trained gender-specific and gender-independent models and compared performance. The gender-specific models may generalize better, since it is well-documented that there are inherent gender differences in how distressed couples express themselves [46]. However, the gender-independent models may benefit from having twice as much training data, since the gender-specific models are only trained on the instances of a single gender.

3.4 Data pre-processing

3.4.1 SNR estimation

Due to the variable acoustic nature of the Couple Therapy corpus, we first set out to estimate the signal-to-noise ratio (SNR) of each session, so we could disregard sessions that were too noisy to analyze. For each session's audio file, we ran a voice activity detector (VAD) that hypothesized whether each 10 ms interval was speech or non-speech. This VAD used a novel long-term signal variability measure, which describes the degree of non-stationarity of the signal, to robustly discriminate speech from silence and background noise [77]. It was specifically designed as a front-end for automatic speech recognition (ASR) and was optimized to detect regions of non-speech longer than 300 ms. We trained the VAD on a 60 s audio clip from one of the held-out sessions with the missing psychology codes (see Section 3.2).

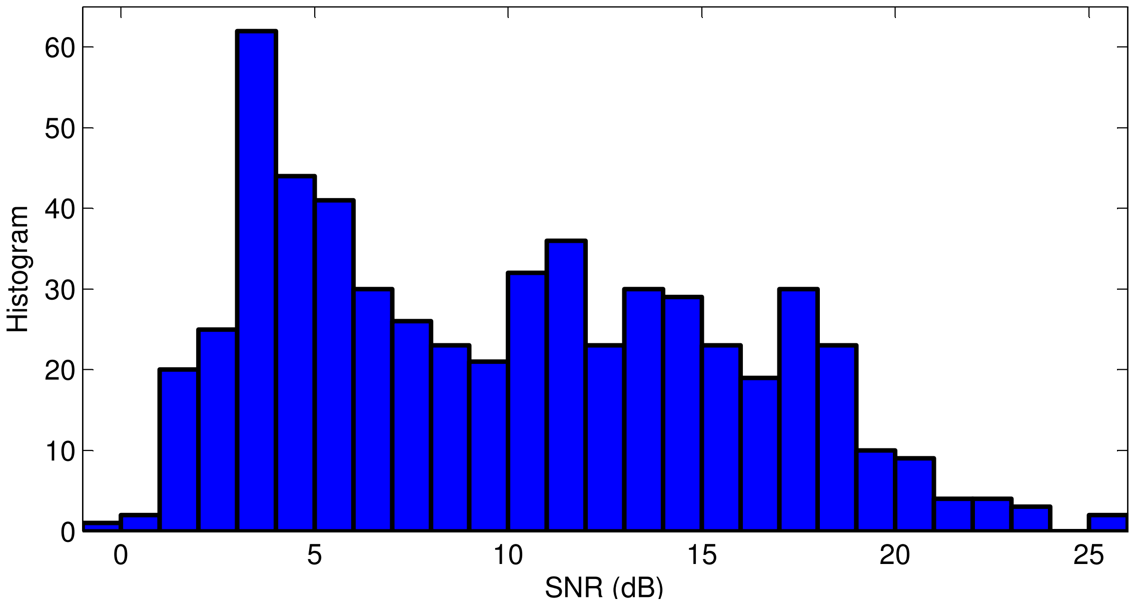

Figure 3.3: A histogram of the estimated average signal-to-noise ratio (SNR) for each of the 569 coded sessions, computed using Equation 3.1.

We used the VAD output to estimate the average SNR of each session's audio file using Equation 3.1, where $\{A_i\} \in S$ is the set of amplitudes endpointed within the speech regions (according to the VAD), and $\{A_i\} \notin S$ is the complement set of amplitudes (deemed to be non-speech by the VAD):

$$\text{SNR (dB)} = 10 \log_{10} \frac{\frac{1}{|i \in S|} \sum\limits_{i \in S} A_i^2}{\frac{1}{|i \notin S|} \sum\limits_{i \notin S} A_i^2} \tag{3.1}$$

Figure 3.3 shows a histogram of the estimated average SNR for the 569 coded sessions. We heuristically decided to only analyze sessions with an average SNR greater than 5 dB. This was done to ensure that the audio features could be reliably extracted. Of the 569 coded sessions, 415 had an average SNR greater than the chosen threshold of 5 dB (72.9%). The other 154 sessions were deemed too noisy for the present work.

3.4.2 Speaker segmentation

Since the Couple Therapy corpus consists of dyadic conversations, we set out to segment the sessions by speaker. This would then allow us to model the interaction appropriately and extract meaningful features for each spouse. In many pattern recognition research involving realistic and complex multi-person interactions, it is common practice to manually segment the data into speaker turns as a pre-processing step. This is typically done for a number of reasons: it ensures that system errors are due to other design factors (e.g., features, learning algorithm); it circumvents the added overhead of implementing automatic segmentation; achieving sufficient performance using automatic methods may be too challenging due to inherent data limitations (e.g., far-field sensors, variable acoustic conditions). However, manually segmenting a corpus of this size was not practical and is not scalable.

In this thesis and in our previous work [18, 109], we took a unique "hybrid" manual/automatic speaker segmentation approach that exploited the available transcriptions with speaker labels. We implemented a recursive automatic speech recognition (ASR)-based procedure to align the transcription with the corresponding audio using *SailAlign* [104], open-source software we developed as part of this work. The iterative algorithm was based on the work by [130], with the extension that aligned portions of the audio were used to adapt the acoustic models at each iteration.

Figure 3.4 is a block diagram of the procedure, showing the flow from the required inputs to the desired output of speaker-segmented audio. Generic acoustic models (AM) and session-specific language models (LM) were used to run ASR on the audio file, aided

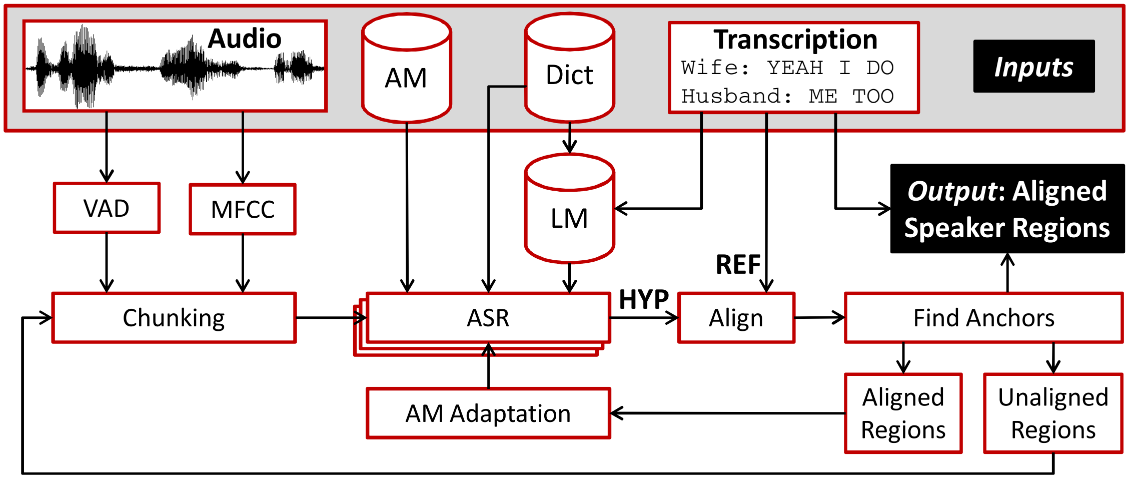

Figure 3.4: Block diagram of the "hybrid" manual/automatic speaker segmentation procedure, implemented using *SailAlign*. See Section 3.4.2 and [104] for details.

by the VAD that split the MFCC feature vector into 15 s chunks. Anchor regions were accepted if aligned portions between the reference transcript (REF) and ASR transcript (HYP) contained at least three consecutive words. The process was then iterated between anchor regions, with AM adaptation at each iteration. Please see [104] for full details on the algorithm.

After *SailAlign* converged, the session was split into wife and husband speaker homogeneous regions and unknown regions in which speech-text alignment could not be achieved (due to multiple factors, including: noisy audio, speaker overlap, and transcription errors). Note that unknown regions that occurred in the middle of a speaker's turn could be merged with the neighboring speaker-homogeneous regions. Figure 3.5 shows that this interpolation-like procedure allowed us to segment 8.7% more words per session, on average, into speaker-homogeneous regions. This figure also shows that we were still not able to align or segment a large percentage of the words in the transcription for some of the 415 sessions that met the 5 dB SNR threshold. For this dissertation, we ignored

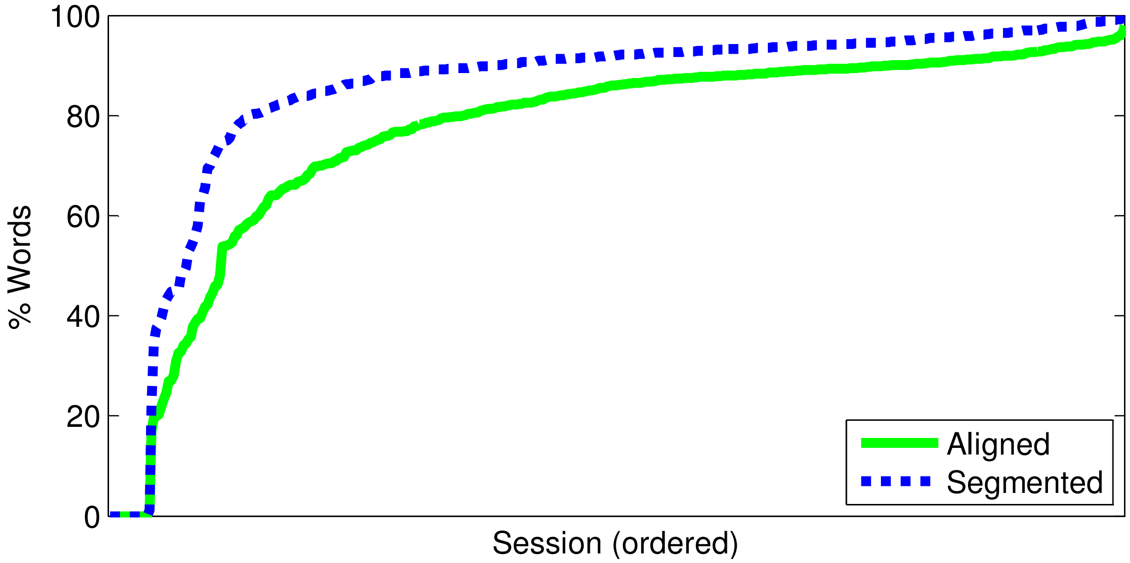

Figure 3.5: The percentage of words *aligned* using *SailAlign* and the percentage of words that were subsequently *segmented* into single speaker regions for the 415 sessions with SNR greater than 5 dB.

the 43 sessions in which we could not segment at least 55% of both the wife's and husband's transcribed words into speaker-homogeneous regions. This left 372 sessions that met both the SNR and speaker segmentation criteria; counting only these sessions, an average of 90.7% of the wives' words and 89.9% of the husbands' words were segmented into speaker-homogeneous regions.

This speaker segmentation procedure provided us hypotheses on when each spouse was speaking, but since we did not have access to the ground-truth times for these speaker turns, we did not have an easy way to evaluate the speaker segmentation performance. One way would be to randomly sample speaker-homogeneous regions and manually verify the speaker. Rather than relying on this laborious method, we instead devised a procedure that exploited the female-male nature of the dyadic interaction participants in this corpus.

The average adult female's speech has a mean fundamental frequency (f_0) of about 210 Hz, while for adult male's speech, it is about 120 Hz [168]. We estimated f_0 for each

Speaker	Mean f_0 (Hz)		Mean *SD* of f_0 (semitones)	
	CT corpus	Traunmüller	CT corpus	Traunmüller
Wife	194	211	3.5	3.4
Husband	121	119	4.0	3.4
Unknown	166	–	5.8	–

Table 3.4: f_0 statistics for the Couple Therapy (CT) corpus, computed across the 3 speaker regions of the 372 sessions, and compared to the female/male statistics listed in [168].

session (see Section 3.5) and computed f_0 statistics for the husband and wife across the speaker-homogeneous regions.

Figure 3.6 shows that there is a clear separation between the mean f_0 values of the wives and husbands (73 Hz on average). In addition, Table 3.4 shows that the average f_0 statistics are similar to the ones reported in [168], computed from hundreds of adult speakers of European languages. This f_0 "sanity check" implies that the speaker segmentation procedure successfully separated the female and male speakers. Importantly, since f_0 is a relatively difficult acoustic cue to track, it also implies that the data quality of the 372 sessions was adequate to robustly extract speech-related audio cues.

In our previous work, in which we used a speech-text alignment procedure without acoustic model adaptation [18], we were only able to achieve a similar level of speaker segmentation performance for 293 sessions. Thus, *SailAlign* enabled us to use 79 more sessions, a relative increase of 27.0%. In total, these 372 sessions are 65.4% of the original 569 coded sessions and total 62.8 hours of data across 104 unique couples.

3.5 Audio feature extraction

With the 372 sessions segmented by speaker, we are now able to extract acoustic features that can be used to predict the six behavioral codes. Spoken cues (e.g., prosody) have been shown to be relevant indicators of a variety of behaviors in the psychology literature (e.g., [51,101]), including in those related to marital interactions [10,83,84]. Affect/emotion are discussed as critical components to communication and are oftentimes conveyed vocally.

In our previous paper [18], we extracted a number of common prosodic/spectral features that have been used in a variety of human-centered engineering tasks, including affect/emotion recognition [86, 111, 113, 148, 152, 153, 181].

We examined an expanded set of features in this thesis by taking an overgenerative approach to feature extraction. This was done for three main reasons: 1) while there is considerable insight in psychology literature on cues that are informative in marital discussions, it is difficult to come up with mappings from these semantic cues to corresponding signal cues, 2) in addition to being informed by psychology, we can also learn from our findings (see Section 3.7, Figure 3.8), and 3) this work represents the first attempt to automatically learn high-level behavioral codes with acoustic features for this corpus. Thus, we explored many common feature types, so a comparison could be made and improved upon in subsequent studies.

In total, we extracted 40,479 session-level features for the gender-specific models and 67,465 session-level features for the gender-independent models. We refer to these as session-level because they describe some aspect of the spouses' behaviors across the entire session. As introduced in Section 3.3.3, the session-level features were computed

Component	Sub-component
LLD	speaking rate, inter-turn pauses, speech/non-speech (VAD), f_0, intensity, 15 MFCCs, 8 MFBs, jitter, jitter-of-jitter, shimmer
Speaker	rated spouse only, partner of rated spouse only, full session, wife only†, husband only†
Granularity	global, halves, hierarchical (hier.) with window durations: 0.1 s, 0.5 s, 1 s, 5 s, 10 s
Functional	mean*, median*, standard deviation*, 1^{st} percentile*, 99^{th} percentile*, $99^{th} - 1^{st}$ percentile*, skewness, kurtosis, minimum position, maximum position, lower quartile, upper quartile, interquartile range, linear approximation slope

Table 3.5: A list of the four components (with sub-components) that make up the session-level features. The starred (*) functionals are the six "basic" functionals. The speaker domains marked with a † are only applicable to the gender-independent models.

as static functionals of low-level descriptors at various temporal granularities over each speaker domain of the session. Therefore, each session-level feature is described by four components: 1) LLD, 2) speaker domain, 3) temporal granularity, and 4) functional. Table 3.5 lists each of these components and Sections 3.5.1-3.5.4 provide further details.

3.5.1 Low-level descriptors

We refer to low-level descriptors (LLDs) as feature streams that are estimated/extracted at fine temporal resolutions (e.g., every 10 ms). Table 3.5 lists each of the LLDs we selected for this dissertation, based on our previous work [18] and on the 2009 Interspeech Emotion Challenge [153] and 2010 Interspeech Paralinguistic Challenge [154].

We computed the mean syllable speaking rate for each aligned word directly from the automatic word alignment results with the help of a syllabified pronunciation dictionary.[2] Therefore, this speaking rate LLD was at the word-level and only applicable to words that were aligned with *SailAlign* (see Section 3.4.2). Another LLD we extracted directly from

[2] http://www.haskins.yale.edu/tada_download/index.php

the alignment results (when available) were the inter-turn durations, measured as the time in seconds from the end of one speaker's turn to the beginning of the next speaker's turn.

We used the VAD speech/non-speech hypotheses to create two LLD vectors: one with the durations of all the speech regions (when the VAD deemed the audio to be speech for consecutive frames), and another with the durations of all the non-speech regions.

We next extracted the following LLDs across each *speech* region every 10 ms using a 25 ms Hamming window: fundamental frequency (f_0), intensity, 15 Mel-frequency cepstral coefficients (MFCCs), 8 log Mel-frequency bands (MFBs), local jitter, jitter-of-jitter (delta jitter), and local shimmer. f_0 and intensity were extracted with Praat [29], and the other LLDs were extracted with openSMILE [72]. The following paragraphs will describe how we computed and normalized these various LLDs, with specific attention paid to f_0 due to the unique characteristics of the Couple Therapy corpus.

Pitch has been shown to be important in affective speech production [101] and emotion recognition research [32,36,37,86,111,181]. f_0 can be estimated from audio and is related to pitch perception. Unfortunately, f_0 is relatively difficult to estimate from speech, since it involves the computation of periodicity from a non-stationary quasi-periodic signal. We used Praat's state-of-the-art autocorrelation function-based f_0 estimator in this research [29]. However, since this is a time-domain approach, it is still susceptible to many common errors.

One of the major types of errors for autocorrelation-based f_0 estimators is pitch halving/doubling [42, 52, 136]. We attempted to minimize these f_0 errors by exploiting the speaker segmentation and using region-specific f_0 range heuristics: 100-400 Hz during wife regions, 70-300 Hz during husband regions, and 70-400 Hz during unknown regions. Therefore, we estimated the f_0 of each session three separate times with the three region-specific ranges and chose the appropriate f_0 estimate based on the speaker segmentation results. The resulting f_0 signal was then passed through an algorithm that attempted to fix instances of pitch halving/doubling by detecting large jumps in the f_0 difference vector.

The f_0 signal was further processed by zeroing it during regions deemed by the VAD to be non-speech and interpolating across unvoiced regions with duration less than 300 ms (using piecewise-cubic Hermite interpolation). We did *not* interpolate across non-speech regions (according to the VAD) or speaker-change points. Finally, the f_0 signal was median-filtered (with a window of length 5) to smooth out any spurious noise; see Figure 3.7 for an example.

Normalization of the raw LLD streams is important, since the final session-level features will be used to train speaker-independent models. We produced two normalized f_0 signals to account for inter-person variations in the mean pitch. The first normalization method, Equation 3.2, subtracts the mean f_0 (μ_{f_0}) of the speaker (wife, husband, or unknown) for each frame. The second method, Equation 3.3, performs a similar transformation on a logarithmic scale, since this may be more perceptually motivated [55].

The μ_{f_0} values were computed across the whole session using the speaker segmentation results; unknown speaker regions were treated as coming from one "unknown speaker."

$$\bar{f}0_{\text{lin}} = f_0 - \mu_{f_0} \tag{3.2}$$

$$\bar{f}0_{\text{log}} = \log_2 \left(\frac{f_0}{\mu_{f_0}} \right) \tag{3.3}$$

The computation of intensity for an audio signal is more straight-forward than estimating f_0. We normalized the intensity LLD to account for differences in microphone levels (caused by variable distances from the microphone to the speakers). Equation 3.4 shows how we normalized each frame-level intensity value, where the μ_{int} values were the mean intensity of the speaker during speech regions, computed across the whole session:

$$\text{int}_n = \frac{\text{int}}{\mu_{\text{int}}} \tag{3.4}$$

We used openSMILE to extract spectral and voice quality features using the same parameter settings as the 2010 Interspeech Paralinguistic Challenge [154]. Short-term spectral features have been successfully used widely in speech processing. We extracted the first 15 MFCCs, computed using the standard bank of 26 triangular filters that were evenly centered along the Mel-frequency scale from 20 Hz to 8000 Hz. To account for environmental and speaker variability, all MFCCs were normalized by performing cepstral-mean subtraction, using Equation 3.5, where the $\mu_{\text{MFCC}[i]}$ values were the mean MFCC of the i^{th} coefficient of the speaker, computed across the whole session:

$$\text{MFCC}_n\left[i\right] = \text{MFCC}\left[i\right] - \mu_{\text{MFCC}[i]}, i = 0, \ldots, 14 \tag{3.5}$$

In addition to these normalized MFCCs, we also filtered the audio with a coarser bank of only 8 triangular filters and computed the log energies at the output. These are the so-called MFB features that are expected to capture coarser spectral characteristics. The filters were evenly centered along the Mel-frequency scale from 20 Hz to 6500 Hz.

Finally, we extracted three voice quality LLDs: local jitter, jitter-of-jitter (delta jitter), and local shimmer. Voice quality attributes have been shown to play a significant role in communicating emotions [79], although most engineering studies have found they are often less discriminative than the more traditional prosodic and spectral features (e.g., [156]), most likely because the uncertainty in estimating the voice quality attributes can overpower the discriminative information they convey.

All three voice quality LLDs are based on the f_0 estimates. Local jitter quantifies period length variations in f_0 and is computed as the average absolute difference between consecutive periods, divided by the average period length of all periods in the frame. Jitter-of-jitter is computed as the average absolute difference between consecutive differences between consecutive periods, divided by the average period length of all periods in the frame. Local shimmer quantifies amplitude variations and is computed as the average absolute difference between the interpolated peak amplitudes of consecutive periods, divided by the average peak amplitude of all periods in the frame [72].

3.5.2 Speaker domains

For all the LLDs described in Section 3.5.1, we extracted features across three separate speaker domains for the gender-specific models and five speaker domains for the gender-independent models. See Table 3.6 for a depiction on which speech regions (wife and/or husband) were included in the various speaker domains.

For the gender-specific models, the three speaker domains were: 1) during speaker-homogeneous regions (according to the speaker segmentation results) where the spouse being *rated* was the speaker (i.e., for the wife-specific models in which the wife was always being rated, the *rated* speaker domain consisted of all the wife speech regions); 2) during speaker-homogeneous regions where the *partner* of the spouse being rated was the speaker; and 3) across the entire session (regardless of speaker).

For the speaker-independent models, we extracted features across five speaker domains: 1) during speaker-homogeneous regions where the spouse being *rated* was the speaker (i.e., for the wife instances, the *rated* speaker domain consisted of the wife speech regions, whereas for the husband instances, the *rated* speaker domain consisted of the husband speech regions); 2) during speaker-homogeneous regions where the *partner* of the spouse being rated was the speaker; 3) across the entire session, regardless of who was speaking or who was being rated; 4) during speaker-homogeneous regions where the *wife* was speaking, regardless of who was being rated; and 5) during speaker-homogeneous regions where the *husband* was speaking, regardless of who was being rated. These final two speaker domain sets were *not* included for the gender-specific models because they would be identical to the rated/partner feature sets and therefore add no information.

Rated Spouse	*Speaker Domain*	*Speech in domain?*	
		Wife	Husband
Wife	Rated spouse	✓	
	Partner		✓
	Full session	✓	✓
	Wife only	✓	
	Husband only		✓
Husband	Rated spouse		✓
	Partner	✓	
	Full session	✓	✓
	Wife only	✓	
	Husband only		✓

Table 3.6: A depiction of which speech regions were included in the five speaker domains, depending on which spouse was being rated.

For example, for the wife-specific models (in which the wife was always being rated for all instances), the "rated" speaker regions are always the same as the "wife" speaker regions, and the "partner" speaker regions are always the same as the "husband" speaker regions; see Table 3.6.

Extracting features for these various speaker domains allowed us to model the behaviors of each spouse and the overall interaction. Modeling individual spouse behavior is particularly important since each spouse was rated separately. However, as shown in Table 3.2, extracting features along the entire session may be just as meaningful, since the two spouse's coded behavior within a given session is often positively correlated.

3.5.3 Temporal granularities

The temporal granularity component of the session-level features refers to the time-scale at which we processed the individual LLDs: 1) global, 2) halves, and 3) hierarchical. The *global* temporal granularity looks at the interaction for a particular speaker domain as a whole entity. Thus, we are viewing each LLD as a representative sample of data, from which we can extract useful “global” features about the speaker/interaction. We only extracted global features in our previous paper [18].

For the *halves* granularity, we split each LLD stream into two halves and computed the difference in functionals (see Section 3.5.4) across the two halves. This temporal granularity attempts to capture gradual changes that may occur as the discussion progresses.

The *hierarchical* temporal granularity splits each LLD stream into disjoint windows of equal duration. Functionals are then computed across each window, and the session-level features are then produced by computing functionals of the functionals; more details are provided in Section 3.5.4. The hierarchical temporal granularity was based on the work by [155] and attempts to capture the variable moment-to-moment changes during the interaction. For this research, we tried window durations of 0.1 s, 0.5 s, 1 s, 5 s, and 10 s. Note that we did not compute hierarchical features for the speaking rate LLD, inter-turn pause LLD, or the two VAD-derived speech/non-speech LLDs, since these LLDs occurred at a longer time scale, which would have resulted in very few samples within each window.

3.5.4 Functionals

For each combination of LLD, speaker domain, and temporal granularity, we produced the final session-level features by computing a series of static functionals. See Table 3.5 for the full list of 14 functionals that we selected. Note that the 1^{st} percentile, 99^{th} percentile, and $99^{th} - 1^{st}$ percentile represent outlier-robust minimum, maximum, and range statistics, respectively. We chose to use these percentiles to account for cases when the functionals were computed over a long period of time, which is particularly relevant for the global features.

We only computed functionals of prosodic and spectral LLDs over *speech* regions (according to the VAD), and we disregarded all zero values (unvoiced regions) when computing the f_0 and voice quality functionals.

For the computation of the hierarchical session-level features, we computed the full 14 functionals over each window. However, to avoid producing an enormous set of session-level features, we only computed six "basic" functionals when computing the functionals-of-functionals; a similar procedure was followed in [155]. These six basic functionals are starred (*) in Table 3.5. In addition, since there was only a limited number of aligned speaker-change points in a session (35.6, on average), we only extracted the six basic functionals for the inter-turn pause LLD.

We also extracted a few dynamic features. For the speech/non-speech (VAD) LLD, we exploited the binary nature of the signal to extract three more session-level features. The first was the probability that a frame was non-speech. We also computed two features based on first-order Markov chain statistics: 1) the probability a frame is non-speech, given that the previous frame was non-speech, and 2) the probability a frame is non-speech, given that the previous frame was speech.

3.6 Prediction of six behavioral codes

Given that there were 372 sessions that were deemed acceptable after pre-processing the corpus (see Section 3.4) and we were only analyzing the top/bottom 20% of the sessions for each spouse/code, we selected the top/bottom 70 sessions for our experiments; the number of unique couples in these 140 selected sessions varied from 68 to 77, depending on the code and rated spouse. With over 40,000 features and only 140 instances for the gender-specific models and over 67,000 features and only 280 instances for the gender-independent models, we became concerned about issues related to dimensionality. However, this type of underdetermined learning scenario (having many more features than instances) is commonplace in genomics and natural language processing problems [98] and emotion recognition [8].

In our previous paper [18], we compared two classifiers: a support vector machine (SVM) with linear kernel, and Fisher's linear discriminant analysis (LDA) with sequential forward feature selection. In this thesis, our initial experiments showed that the LDA did not perform as well, most likely due to the high dimensionality of the feature space and the greedy feature selection method.

In this thesis, we again used linear classifiers since the dimensionality of the feature space (40,000+) was orders of magnitude greater than the number of instances (140-280). We compared four binary linear classifiers: l^2-regularized SVM with linear kernel (SVM-l^2), l^1-regularized SVM with linear kernel (SVM-l^1), l^2-regularized logistic regression (LR-l^2), and l^1-regularized logistic regression (LR-l^1).

The loss functions of the four classifiers, used to find the optimal weight coefficients, are written in Equations 3.6-3.9, where m is the number of training instances, $y_i \in \{-1, 1\}$ is the class label (low/high) for instance i, $\mathbf{x}_i \in R^n$ is the corresponding n-dimensional feature vector, $\mathbf{w} \in R^n$ is the linear weight vector, $\|\mathbf{w}\|_1$ is the l^1-norm of $\mathbf{w}$, and C is a tuning penalty parameter ($C > 0$).

While the l^2-regularized versions of the classifiers (Equations 3.6 and 3.8) are more commonly used, the l^1-regularized classifiers (Equations 3.7 and 3.9) are appealing since they find a sparse solution (some of the weight coefficients will be identically zero). This may be advantageous for two reasons: 1) there are potentially many irrelevant and redundant features due to the overgenerative nature of the feature extraction process (see Section 3.5), so dimensionality reduction via sparse solutions may lead to more robust estimates of the weight coefficients and improved classification, and 2) sparse solutions are more interpretable and provide a means to determine the relative importance of the features.

$$\hat{\mathbf{w}}_{\text{SVM-}l^2} = \min_{\mathbf{w}} \left(\tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m} \max\left(0, 1 - y_i \mathbf{w}^T\mathbf{x}_i\right)^2 \right) \tag{3.6}$$

$$\hat{\mathbf{w}}_{\text{SVM-}l^1} = \min_{\mathbf{w}} \left(\|\mathbf{w}\|_1 + C\sum_{i=1}^{m} \max\left(0, 1 - y_i \mathbf{w}^T\mathbf{x}_i\right)^2 \right) \tag{3.7}$$

$$\hat{\mathbf{w}}_{\text{LR-}l^2} = \min_{\mathbf{w}} \left(\tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m} \log\left(1 + e^{-y_i \mathbf{w}^T\mathbf{x}_i}\right) \right) \tag{3.8}$$

$$\hat{\mathbf{w}}_{\text{LR-}l^1} = \min_{\mathbf{w}} \left(\|\mathbf{w}\|_1 + C\sum_{i=1}^{m} \log\left(1 + e^{-y_i \mathbf{w}^T\mathbf{x}_i}\right) \right) \tag{3.9}$$

We used the implementations in LIBLINEAR for all four classifiers [73]. Note that the primal forms of the loss functions are written in Equations 3.6-3.9 for clarity. In practice, the dual forms were faster to train; see [73] for details.

Prior to training the classifiers, we z-normalized all features at each cross-validation fold by subtracting the mean value in the training set and dividing by the standard deviation. This feature scaling was done to ensure that the regularization would be applied evenly to all features. As mentioned in Section 3.3.3, the tuning parameter C

was optimized for each classifier at each train/test cross-validation fold by using a grid search and choosing the value with the highest average classification accuracy on the training set using 5-fold couple-disjoint cross-validation.

For all four classifier implementations, we generated a class hypothesis ($\hat{y}$) on a test instance by taking the sign of the inner product between the optimal weight vector ($\hat{\mathbf{w}}$) and the feature vector ($\mathbf{x}$) of the test instance:

$$\hat{y} = \operatorname{sgn}\left(\hat{\mathbf{w}}^T\mathbf{x}\right) \tag{3.10}$$

3.7 Results and Discussion

Table 3.7 displays the results for the wife and husband instances for all six codes, both model types (gender-specific and gender-independent), and all four classification methods (SVM-l^2, SVM-l^1, LR-l^2, and LR-l^1). These results are compared to the baseline chance performance of 50% accuracy and the upper-bound in performance as computed from the individual human evaluator scores (Table 3.3). We see from Table 3.7 that the classification performance ranged from below chance accuracy (49.3% for the husband-specific SVM-l^1 classifier for sadness) to as high as 85.7% (for husband's global negative affect). Performance varied greatly as a function of the various factors (spouse being rated, model type, classifier, and code). In this section, we provide statistical analyses to compare these various factors; Section 3.9 discusses ongoing and future work to improve upon the results achieved in this thesis.

Model	Classifier	*acc*	*bla*	*pos*	*neg*	*sad*	*hum*	AVG
Wife is the spouse being rated (140 instances total)								
Baseline	Chance	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Gender-Specific	SVM-l^2	75.0	**85.0**	74.3	79.3	**67.9**	**67.1**	74.8
	SVM-l^1	75.7	81.4	73.6	75.7	56.4	57.9	70.1
	LR-l^2	**77.9**	84.3	74.3	**80.0**	66.4	**67.1**	**75.0**
	LR-l^1	72.9	80.7	**77.9**	77.9	55.7	59.3	70.7
Gender-Indep.	SVM-l^2	75.0	82.9	74.3	78.6	63.6	64.3	73.1
	SVM-l^1	75.0	80.7	72.9	76.4	52.9	52.1	68.3
	LR-l^2	**77.9**	82.1	75.7	**80.0**	62.9	65.0	73.9
	LR-l^1	76.4	80.7	72.1	77.1	60.0	57.9	70.7
Up-Bound	Human	96.7	99.6	98.5	98.6	93.9	96.5	97.3
Husband is the spouse being rated (140 instances total)								
Baseline	Chance	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Gender-Specific	SVM-l^2	**78.6**	72.9	72.1	84.3	57.9	69.3	72.5
	SVM-l^1	67.1	73.6	67.9	**85.7**	49.3	63.6	67.9
	LR-l^2	**78.6**	72.9	72.1	84.3	**60.0**	**71.4**	**73.2**
	LR-l^1	77.1	75.0	71.4	85.0	52.9	64.3	71.0
Gender-Indep.	SVM-l^2	**78.6**	75.7	**72.9**	**85.7**	**60.0**	63.6	72.7
	SVM-l^1	72.1	**80.7**	69.3	81.4	57.9	56.4	69.6
	LR-l^2	77.1	75.7	72.1	85.0	59.3	68.6	73.0
	LR-l^1	75.7	77.9	67.9	83.6	57.9	65.0	71.3
Up-Bound	Human	96.7	98.1	97.4	98.0	84.9	97.1	95.4

Table 3.7: Percentage of correctly classified instances for the wives and husbands, 6 codes, 2 model types (gender-specific and gender-independent), and 4 classifiers: support vector machine (SVM) and logistic regression (LR), with l^2 and l^1 regularization (Equations 3.6-3.9). Baseline chance performance was 50%, and an upper-bound was estimated using individual human evaluator scores (Section 3.3.2).

We used two statistical tests to determine if the differences in performance were statistically significant: a one-sided McNemar's test for "paired" instances [122], which occurred when comparing results from the same code and same rated spouse gender (wife or husband); and a one-sided difference in binomial proportions test for non-paired instances [123], which occurred when comparing different codes or different rated spouse genders.

All results shown in Table 3.7 were significantly better than the chance baseline at the 5% significance level, except the following: all l^1-regularized classifiers for wife's sadness and use of humor, all classifiers for husband's sadness, and the gender-independent l^1-regularized SVM classifier for husband's use of humor. All classifiers performed significantly worse than the estimated upper-bounds (all $p < 0.001$).

In our previous paper [18], in which we only used "global" features and analyzed 100 wife and husband instances per code, we achieved an average classification accuracy of 70.15% (averaging across the six codes and both rated spouse genders). In this dissertation, we analyzed 140 wife and husband instances per code, and the average classification accuracy for the best overall system (gender-specific LR-l^2) was 74.1%, an absolute improvement of 3.95% and a relative improvement of 5.63%. This difference in performance is significant ($p < 0.01$), so this extended research effort has helped reduce the gap between automatic and human coders for this particular behavioral coding problem.

We see from Table 3.7 that for most cases, the gender-specific models outperformed the gender-independent models, the l^2 classifiers outperformed the l^1 classifiers, and logistic regression outperformed the SVM classifiers. For both the husband and wife instances, the best overall system with the highest average code performance was trained in a gender-specific manner with l^2-regularized logistic regression. For the wife instances, this best average code performance (75.0%) was significantly higher than all four l^1 classifiers (all $p < 0.05$) but was not significantly higher than the other three l^2 classifiers. For the husband instances, this best average code performance (73.2%) was only significantly higher than the gender-specific SVM-l^1 classifier ($p < 0.01$).

Feature Subset (f)		N_f		$N_{f,sel}$			$N_{f,sel}/N_{sel}$			$N_{f,sel}/N_f$		
component	sub-comp.	W&H	I	W	H	I	W	H	I	W	H	I
(all)	(all)	40479	67465	896.8	1051	1322	1.000	1.000	1.000	0.022	0.026	0.020
LLD	rate	48	80	2.838	3.218	3.512	0.003	0.003	0.006	0.059	0.067	0.044
	VAD/turn	111	185	10.00	9.275	12.79	0.012	0.009	0.018	0.090	0.084	0.069
	f_0	4032	6720	114.2	116.3	151.0	0.160	0.125	0.165	0.028	0.029	0.023
	int_n	1344	2240	25.34	29.68	38.13	0.024	0.027	0.021	0.019	0.022	0.017
	MFCC_n	20160	33600	450.4	543.7	678.4	0.493	0.519	0.496	0.022	0.027	0.020
	MFB	10752	17920	192.2	214.5	282.8	0.210	0.192	0.194	0.018	0.020	0.016
	V.Q.	4032	6720	101.8	134.5	155.0	0.098	0.124	0.101	0.025	0.033	0.023
Speaker	rated	13493	13493	310.0	399.7	303.2	0.351	0.392	0.219	0.023	0.030	0.023
	partner	13493	13493	291.4	333.6	273.4	0.315	0.306	0.169	0.022	0.025	0.020
	both	13493	13493	295.5	317.9	296.2	0.334	0.302	0.229	0.022	0.024	0.022
	wife	—	13493	—	—	223.8	—	—	0.173	—	—	0.017
	husband	—	13493	—	—	225.1	—	—	0.210	—	—	0.017
Granularity	global	1419	2365	38.71	41.31	52.68	0.046	0.039	0.056	0.027	0.029	0.022
	halves	1260	2100	36.75	43.86	52.75	0.044	0.052	0.047	0.029	0.035	0.025
	hier.–all	37800	63000	821.3	966.1	1216	0.910	0.909	0.897	0.022	0.026	0.019
	hier.–0.1s	7560	12600	132.4	164.5	206.7	0.154	0.169	0.162	0.018	0.022	0.016
	hier.–0.5s	7560	12600	147.7	170.2	220.8	0.154	0.152	0.156	0.020	0.023	0.018
	hier.–1s	7560	12600	165.0	197.6	251.1	0.179	0.189	0.188	0.022	0.026	0.020
	hier.–5s	7560	12600	188.4	218.5	271.5	0.215	0.198	0.202	0.025	0.029	0.022
	hier.–10s	7560	12600	187.8	215.2	266.1	0.208	0.201	0.189	0.025	0.029	0.021

Table 3.8: For each feature subset (f), we show the total number of features (N_f), the number of selected features ($N_{f,sel}$), the fraction of the selected features that were from a given feature subset ($N_{f,sel}/N_{sel}$), and the probability of a feature being selected for a particular feature subset ($N_{f,sel}/N_f$). For clarity, we only displayed mean results for the l^1-regularized logistic regression classifier, averaged across all codes and cross-validations. Results are shown for the wife-specific (W), husband-specific (H), and gender-independent (I) models. Note that the "wife" and "husband" speaker domain feature subsets were not included in the gender-specific models (see Section 3.5.2).

Overall, the advantages of using the l^1 classifiers (sparse solutions that are easier to interpret) did not lead to higher classification performance. One possible explanation for the relatively poor performance for many of the l^1-regularized classifiers may be that the selected features did not generalize well. Also, the l^1 cost functions (Equations 3.7 and 3.9) may be more difficult to optimize, with classification performance being more sensitive to the selection of the tuning parameter (C).

While most of the performance differences between the logistic regression and SVM classifiers were not significant, the logistic regression models had higher overall code performance. While SVMs are considered to be a state-of-the-art binary classifier, logistic regression has the advantage of being a simpler model to train.

There were few significant differences between the gender-specific and gender independent models, with the gender-specific models performing better on average for both the wife and husband instances. One possible explanation for this difference can be explained in the psychology literature, which says that women and men express themselves differently [46]; this implies that gender-specific classifiers are more appropriate. We can also provide a more data-driven explanation: the advantage of having twice as much training data for the gender-independent models was not as important as the advantage of having features that were gender-matched (as in the gender-specific models). The gender-independent models could most likely be improved in the future by normalizing the acoustic features by speaker *and* gender.

Comparing between codes (for the best performing classifiers/models only), we found for the wife instances: performance in classifying sadness and humor was significantly lower than classifying acceptance (both $p<0.05$), blame (both $p<0.001$), global positive affect (both $p < 0.05$), and global negative affect (both $p < 0.01$). For the husband instances, classification performance for sadness was significantly worse than all other codes (all $p < 0.05$). In addition, performance in classifying blame was significantly higher than humor ($p < 0.05$), and global negative affect was significantly higher than global positive affect and humor (both $p < 0.005$). The large range in classification performance may be due in part to the fact that some codes are inherently more difficult to separate; see Figure 3.1, where small separations between low/high code scores (e.g., husband's sadness) implies that the extreme behaviors are perceptually closer.

There were no significant differences when comparing the classification performance for the wife instances versus the husband instances for the 6 codes or the average code performance (using the best performing classifiers/models). The higher average code performance for the wife instances over the husband instances for the gender-specific LR-l^2 classifier (1.8% difference) may be partially explained by the upper-bounds in automatic performance. We see in Table 3.3 that the average upper-bound performance for the wife instances (97.3%) was 1.9% higher than the average upper-bound performance for the husband instances (95.4%). This indicates that the human evaluators tended to agree more often when rating the wive's behaviors for the six codes, which suggests that automatically classifying the husband's behavior may be more difficult for this subset of data.

To compare the relative importance of the various features, we analyzed which features had non-zero weight coefficients for the l^1-regularized logistic regression classifiers at each train/test fold. We refer to these features as the "selected features." Table 3.8 provides details on the number and fraction of selected features for the various feature components and sub-components ("subsets") proposed in this dissertation (see Section 3.5).

In the first row of Table 3.8, we see that on average, approximately 1090 features were selected by the LR-l^1 classifiers at each train/test fold. This represents about 2.3% of the full feature set, which is a significant reduction in the dimensionality of the feature space. We also see in Table 3.8 that the number of selected features for a given feature subset was proportional to the dimensionality of the subset, as demonstrated in the $N_{f,sel}/N_{sel}$ columns; for example, only approximately 0.3% of the selected features were *rate* LLD features, whereas approximately 91% of the selected features were *hierarchical* temporal granularity features.

An interesting finding is in the $N_{f,sel}/N_f$ columns of Table 3.8, which show the probability that a feature is selected for a given feature subset. With the exception of the *rate* and *VAD/turn* LLD subsets (which have a relatively low dimensionality), all proposed feature subsets had a similar probability (ranging from 0.018-0.035 for the gender-specific models and 0.016-0.025 for the gender-independent models). This implies that all proposed feature subsets contain relevant information for learning the behavioral codes but that this information may be spread across a larger number of features for the higher dimensional feature subsets. For the rate and VAD/turn LLD subsets, the relatively large probability of a feature being selected may be due to the fact that there is less redundancy in these low-dimensional feature subsets.

We ran a second set of classification experiments using single feature subsets, so we could empirically compare their relative performance in predicting the behavioral codes. For these experiments, we only trained gender-specific models using l^2-regularized logistic regression. Figure 3.8 is a bar plot of these results, where we also show the performance for the case when we trained the classifier on all features.

We see in Figure 3.8 that all feature subsets performed better than chance (50%). Also, we achieved the best average code classification performance with single feature subsets using the *MFCC* and f_0 LLDs, the *rated* and *both* speaker domains, and the *global* and *hierarchical* temporal granularities. This suggests that these features may be the most relevant for this automatic behavioral coding problem.

Importantly, we attained the highest average code classification performance when using all features, with one exception: for the husband instances, we achieved the highest accuracy when using only *rated* speaker domain features. This means that for the husband instances, the *partner* and *both* speaker domain features did not help improve the average code classification performance. This is not an unreasonable finding since the husband was the person being rated, and it suggests that the husband's speech regions are most informative for predicting the husband's behavioral code scores. On the other hand, for the wife instances, we found that the *both* speaker domain features were best, which implies that features derived from the entire interaction were most informative for classifying the wife's behaviors.

The previous experiments demonstrated the utility of the proposed acoustic features and classifiers in discriminating *extreme* behaviors (top/bottom 20%) from the spouses. However, quantifying the less extreme instances that fall in the middle 60% of the code range is also crucial to automate a behavioral coding system. To provide insight into how well we can quantify these middle instances, we performed one final experiment.

We first trained the binary gender-specific l^2-regularized logistic regression classifier as before, in a leave-one-couple-out manner using the top/bottom 20% of the data. We then applied the trained model to the remaining 60% of the data to attain posterior probability estimates (i.e., the probability of belonging to the "high" code class and the "low" code class). Our hypothesis is that instances with *higher* evaluator code scores will have *higher* "high" code posteriors than instances with *lower* evaluator code scores. To test this hypothesis, we computed the Spearman's rank correlation coefficient between the "high" code posteriors and the mean evaluator scores. Spearman's correlation was

Rated-System	*acc*	*bla*	*pos*	*neg*	*sad*	*hum*	AVG
Wife-auto	0.24	0.21	**0.34**	**0.36**	**0.35**	0.16	**0.28**
Wife-eval	**0.36**	**0.45**	**0.28**	**0.41**	0.10	**0.40**	**0.33**
Husband-auto	0.12	0.14	**0.28**	0.22	0.22	0.10	0.18
Husband-eval	0.19	**0.38**	**0.26**	**0.32**	0.03	**0.40**	**0.26**

Table 3.9: Performance in ranking the instances of the middle 60% of the six behavioral codes for the automatic (auto) system, as compared to inter-evaluator agreement (eval). Shown here is the mean Spearman's correlation, with the bold numbers significant at the 5% level.

the chosen metric because it compares the relative order of the instances, *not* the actual values themselves.

To establish how difficult it is for humans to rank these middle instances, we computed inter-evaluator agreement for the middle 60% of the code range by randomly sampling individual evaluator's scores from each instance and computing the Spearman's correlation with the mean scores of the other evaluators. We repeated this random sampling procedure 10,000 times.

Table 3.9 and Figure 3.9 show the results from this final experiment for all six codes and both spouses. We see from Table 3.9 that the mean correlations for the automatic system were all positive, which means the binary classifiers trained on the extreme instances were able to rank the middle 60% of the instances better than chance (correlation = 0). The relatively low correlation values for both the automatic system and the evaluators demonstrate the inherent difficulty in quantifying these more neutral/ambiguous behavioral displays of the spouses.

The automatic system performed better at ranking the wife's middle 60%, as opposed to the husband's, a trend also seen with the human evaluators. This implies that men's less extreme behavior may have a relatively higher degree of variability and/or ambiguity, as compared to women's.

We also see in Table 3.9 and Figure 3.9 that, on average, the inter-evaluator correlation was higher than the automatic performance, which was expected; 11 out of the 14 *evaluator* correlations listed in Table 3.9 were significant at the 5% level, while only 5 out of the 14 *automatic* correlations were significant. However, there were cases when the automatic system ranked the instances better than held-out evaluators (e.g., sadness). This suggests that the automatic system was able to model the average evaluator perception, despite a large degree of individual evaluator variability.

3.8 Fusion of speech and language information

One of the weaknesses with the proposed method in Section 3.6 is that it ignores important lexical cues. In this section, we concentrate on predicting a single code, the spouse's, "level of blame," by fusing automatically-derived speech and language information [16]. The coding manual used to rate blame says that, "explicit blaming statements (e.g., 'you made me do it') warrant a high blame score [93]," and the acoustic features we extracted in Section 3.6 were not able to capture these types of spoken phenomena. Blame is particularly relevant for this domain of data and is oftentimes targeted in couple therapy, since blaming behavior can lead to an escalation of negative affect and resentment between the spouses [60].

In this section, we introduce an ASR-derived classification method that incorporates lexical information through the use of two competitive maximum likelihood language models (one trained on “low blame” text and the other trained on “high blame” text). We show that even with noisy ASR, this method is able to capture discriminative aspects of blame behaviors. Moreover, we show that we can attain the highest classification performance by combining the complementary acoustic and language information sources through score-level fusion of the two classification methods. As part of this work, we also provide an upper bound on performance by running an oracle experiment for the case when we have access to perfect word-level transcriptions.

We used the same binary classification set-up as before by partitioning the data into two classes: *high* blame and *low* blame. The high blame partition consisted of the 70 sessions (approximately 20% of the 372 sessions) with the highest average blame score for the wife and the 70 sessions with the highest average blame score for the husband. The low blame partitions consisted of the 140 sessions with the lowest average blame score: 70 for the wife and 70 for the husband. The blame scores for the two classes ranged from 1.0-1.5 for low blame and 5.0-9.0 for high blame, so they were separable to the human evaluators.

In this section, we chose to train gender-independent models, thus effectively doubling the amount of training data. We chose *accuracy* to be the performance metric, defined as the percentage of correctly classified test sessions (out of 280); baseline chance accuracy is 50%. To ensure that the reported results were not overstated, we used leave-one-couple-out cross-validation to separate training and test data, and we optimized all classifier parameters at each train/test fold by using leave-two-couples-out cross-validation on the training data. Therefore, there was no “contamination” of the test couple during the training stages.

Section 3.8.1 discusses the acoustic feature extraction process for this fusion work. Sections 3.8.2-3.8.5 describe the static acoustic, ASR-derived lexical, oracle lexical, and fusion classifiers, respectively. Figure 3.10 is a block diagram for the signal-driven classification methods.

3.8.1 Acoustic Feature Extraction

For the acoustic classifier, we used the same acoustic speech features listed in Table 3.5. The lexical classification method we implemented (Section 3.8.3) is based on ASR within the hidden Markov model framework. We used the standard frame-level 39-dimensional vector: the first 13 mean-subtracted Mel-frequency cepstral coefficients (MFCCs) and their first-order derivative (Δ) and acceleration ($\Delta\Delta$) coefficients.

3.8.2 Static Acoustic Classifier

The static acoustic classifier is very similar to the one proposed in Section 3.6. It finds a mapping from the high-dimensional static acoustic feature space, which represent various properties of the spouses' speech, to the binary blame class labels. We used the support vector machine (SVM) implementation in LIBSVM [41]. Since there were orders of magnitude more features (50,000+) than instances (280), we used a linear kernel. All features were z-normalized by subtracting the mean value in the training data and dividing by the standard deviation.

3.8.3 ASR-derived Lexical Classifier

The main problem in using ASR to derive lexical information is the resulting "noisy" word hypotheses, due to numerous factors (e.g., noisy audio, mismatched acoustic/language models). We partially circumvented this noisy ASR problem by implementing an ASR-derived lexical classifier, which incorporated differences in language use between *low* and *high* blame spouses via competitive language models. We derive the equation for this classifier in Equations 3.11-3.17, based on [71].

$$\begin{aligned}
[B^*, W^*] &= \operatorname*{argmax}_{B,W} \ P(B, W|O), \quad B \in \{-1, 1\} && (3.11)\\
&\approx \operatorname*{argmax}_{B,W} \prod_t P(B, W_t|O_t) && (3.12)\\
&= \operatorname*{argmax}_{B,W} \prod_t P(O_t|W_t, B)P(W_t|B) && (3.13)\\
&\approx \operatorname*{argmax}_{B,W} \prod_t P(O_t|W_t)\tilde{P}(W_t|B) && (3.14)
\end{aligned}$$

Equation 3.11 states to choose the most probable blame class $B \in \{-1, 1\}$ (low/high blame) and most likely word sequence W, given the acoustic observations O of the rated spouse's speech; we disregard the speech regions of the rated spouse's partner for this classification implementation. For computational reasons, we assume in Equation 3.12 that each speaker turn is independent, and we denote the acoustic observations and word sequence of turn t as O_t and W_t, respectively. We attain Equation 3.13 by applying Bayes' theorem and dropping the B prior, since both blame classes are equally represented in our experiments. Equation 3.13 is a variation of the fundamental equation for ASR, where $P(O_t|W_t, B)$ corresponds to the "blame class"-specific acoustic model (AM), and $P(W_t|B)$ corresponds to the "blame class"-specific language model (LM).

For this initial work, we did not train AMs for both blame classes and instead used generic AMs; thus, we assumed that the acoustic observations were independent from B, as shown in Equation 3.14. We trained the "blame class"-specific LMs using the transcriptions of spouses in the training data at each cross-validation fold: a "high blame" LM on the text from spouses rated as having high blame and a "low blame" LM on the text from spouses rated as having low blame. We trained the LMs on unigram word frequency counts for simplicity and to avoid more complex smoothing procedures and data sparsity issues. Both LMs were smoothed via interpolation with a λ-weighted background (BG) LM trained on out-of-domain text:

$$\tilde{P}(W_t|B) = (1-\lambda)P(W_t|B) + \lambda P(W_t|\mathrm{BG}), \quad 0 < \lambda < 1 \tag{3.15}$$

Since estimating the probability of the most likely path through the ASR word lattice may not be robust, we incorporated the probabilities of the 100 most likely ("N-best") paths through the lattice for each speaker turn. We assumed in our implementation that the N-best hypotheses were independent; see Equation 3.16, where the n subscript refers to the n^{th} most likely path. In practice, we applied Equation 3.17 for numerical reasons. See Figure 3.10 for a depiction of the ASR-derived lexical classifier, where we denote the smoothed LMs for low and high blame as LM_{lo} and LM_{hi}, respectively.

$$
\begin{aligned}
B^* &= \operatorname*{argmax}_{B,W} \prod_n \prod_t P(O_t|W_{t,n})\tilde{P}(W_{t,n}|B) && (3.16)\\
&= \operatorname*{argmax}_{B,W} \sum_n \sum_t \log P(O_t|W_{t,n})\tilde{P}(W_{t,n}|B) && (3.17)
\end{aligned}
$$

3.8.4 Oracle Lexical Classifier

To find an upper bound on the performance of the proposed ASR-derived lexical classifier, we ran an oracle experiment that assumed we had perfect word recognition rate (i.e., we used the manual transcription). This oracle classifier is shown in Equation 3.18, where W is the sequence of transcribed words across the session for the rated spouse, and we used the same smoothed LMs as in Section 3.8.3 to compute $\tilde{P}(W|B)$.

$$
B^* = \operatorname*{argmax}_{B} \tilde{P}(W|B), \quad B \in \{-1, 1\} \qquad (3.18)
$$

3.8.5 Fusion Classifier

Fusion of multimodal information has been advantageously applied in many engineering research domains. For example, improved emotion recognition has been reported when fusing audio/language/discourse features [113] and audio/video features [180]. Fusion typically takes place at the feature-level (e.g., by combining features at the input of a classifier), score-level (e.g., by combining output confidence scores from many classifiers), or decision-level (e.g., by voting on multiple classifier decisions). For our experiments, fusion at the score-level was most applicable, given the high dimensionality of the static acoustic classifier (not ideal for feature-level fusion) and since we only had two classifiers (not ideal for decision-level fusion).

The fusion features FF were computed using Equation 3.19, where conf_c is a non-negative confidence score for classifier c:

$$FF_c = (\text{conf}_c)(B_c^*), \quad B^* \in \{-1, 1\}\,,\ \text{conf} \geq 0 \tag{3.19}$$

For the ASR-derived and oracle lexical classifiers, the magnitude of the difference in log-probabilities between the competing LMs served as the confidence score. For the static acoustic SVM classifier, class probability estimates (made by LIBSVM using internal cross-validation on the training data) were the confidence scores [41].

We again used LIBSVM's SVM for the fusion classifier and z-normalized the fusion features, so they were on a comparable scale. We tried three pairs of classifier combinations: fusing the static acoustic and ASR-derived lexical classifiers (see Figure 3.10), fusing the static acoustic and oracle lexical classifiers, and fusing the two lexical classifiers.

3.8.6 Results

Table 3.10 shows the performance of the various classifiers on the 280 instances. Using a difference in binomial proportions statistical test, we see that all proposed classifiers had significantly higher accuracy than chance accuracy of 50% (all $p < 0.01$). All oracle classifiers had significantly higher accuracy than all non-oracle classifiers (all $p < 0.01$), with no statistical difference between any of the oracle classifiers (all $p > 0.05$). There was no statistically significant difference between any of the non-oracle classifiers ($p >$

0.05), except the acoustic and ASR-derived lexical fusion classifier had significantly higher accuracy than the ASR-derived lexical classifier alone ($p < 0.05$).

In isolation, the oracle lexical classifier (which uses the perfect transcription) performed best, which suggests that lexical information is critical for classifying blame behaviors; this agrees with both intuition and the coding manual [93]. Even though the static acoustic classifier ignores these important lexical cues, it outperformed the ASR-derived lexical classifier, although not significantly ($p > 0.05$). Achieving 75% classification accuracy with the ASR-derived lexical classifier is a promising result, especially considering the noisy acoustic conditions and spontaneous nature of the corpus.

The significant difference between the ASR-derived and oracle lexical classifiers can most likely be attributed to the quality of the ASR word lattices. We found the ASR word error rate ranged from 40%-90% across the sessions (using standard metrics on the most likely word hypothesis). For less noisy data, we would expect the quality of the ASR lattices to improve and the classification performance to increase.

We see from the fusion experiments that performance decreased when we fused the two lexical classifiers, most likely because both of these classifiers model the language use of the spouses. We got a 0.7% absolute (0.8% relative) improvement when we fused the static acoustic classifier with the oracle lexical classifier. Although this difference is not significant, it suggests that the system was able to incorporate complementary acoustic information from the spouses' speech.

Although it is not a statistically significant difference in performance ($p > 0.05$), we saw a 2.5% absolute (3.1% relative) boost in performance when we fused the static

System	*Classifier*	*Acc* (%)
Baseline	Chance	50.0
Unimodal	Acoustic	**79.6**
	Lexical/ASR	**75.4**
	Lexical/Oracle	91.1
Fusion	Acoustic + Lexical/ASR	**82.1**
	Acoustic + Lexical/Oracle	91.8
	Lexical/ASR + Lexical/Oracle	87.5

Table 3.10: The accuracy of the proposed classification methods.

acoustic and ASR-derived lexical classifiers. This fusion classifier advantageously combined automatically derived blaming cues from the spouses' speech and language. It also has the benefit of incorporating confidence scores, which can be interpreted to determine the relative importance of "what the spouse said" versus "how the spouse spoke," with respect to the perception of blame.

3.9 Conclusions

In this chapter, we proposed an engineering methodology toward automating a manual human behavioral coding system for marital problem-solving discussions using acoustic speech features. One of the unique aspects of this research is that we used interaction data from real couples, collected as part of a longitudinal psychology study on couple therapy, and coded with the guidance of expert psychologists. While automatically predicting the spouses' behavioral codes is a challenging problem, developing tools and algorithms that can model complex human behaviors during realistic interactions are one of the main goals in behavioral signal processing (BSP).

After eliminating a third of the audio data because of extreme noise conditions or poor speaker segmentation, we extracted multiple acoustic low-level descriptors and computed static functionals at various temporal granularities to capture global speech properties for both spouses. The resulting high-dimensional feature set was then used to automatically classify the top/bottom 20% of the instances for six selected behavioral codes.

We attained the highest average code classification performance (75% accuracy for the instances when the wife was being rated and 73% accuracy for the instances when the husband was being rated) using l^2-regularized logistic regression; these best models were trained in a gender-specific fashion, with the wife and husband models trained separately. The best code classification performance for the wife instances ranged from 67% for humor to 85% for blame, while the best code classification performance for the husband instances ranged from 60% for sadness to 86% for negativity.

As part of this work, we provided analysis about the relative importance of the various feature subsets we extracted, based on the gender-specific l^1-regularized logistic regression models. We showed that while the higher-dimensional feature subsets made up a larger portion of the "selected" features (features with non-zero weight coefficients), the probability that a feature was selected was similar across all proposed feature subsets. Future work will further investigate dimensionality reduction and feature selection techniques (e.g., [8]) to help find a lower-dimensional and code-specific feature space that can discriminate between the low and high behavioral code scores.

This initial study has led to a number of ongoing research efforts. In addition to computing static functionals across various temporal granularities within the session, we are also experimenting with ways to dynamically model the interaction. Our related and current work has modeled the trajectories of prosodic features to quantify acoustic entrainment effects between the two spouses [109, 110].

We also showed we could successfully separate 82% of the extreme instances of blaming behavior conveyed by the spouses through fusion of automatically derived speech and language information [16]. In the future, we will work to improve: the static acoustic classifier (e.g., by implementing feature selection techniques); ASR-based lexical classifier (e.g., by training "blame class"-dependent acoustic models and experimenting with other procedures to merge N-best hypotheses); and fusion classifier (e.g., by experimenting with new confidence score estimation schemes). See [76] for more related work on using both the manual transcriptions and automatically-generated transcriptions (through automatic speech recognition) to predict the session-level behavioral codes. Alternative acoustic/lexical fusion methods were also considered in [105]. As part of our future work, we also plan to extend these fusion experiments to other behavioral codes.

In addition, since certain portions of the ten-minute discussions may be more relevant than others, we are working on detecting code-specific salient regions. Concurrent work has viewed the automatic classification of the behavioral codes as a multiple instance learning problem. Initial classification experiments that applied the Diverse Density Support Vector Machine framework with both transcription and acoustic features have been promising and allow for the estimation of salient regions during the interaction [78, 105].

We are currently in the process of coding a subset of the Couple Therapy corpus at a finer-grained (continuous) level. This will enable us to use supervised learning techniques to automatically locate the more relevant temporal regions of the interaction. We believe that incorporating saliency detection in an informed manner could allow us to automatically model the interactions in a fashion that more closely resembles trained human evaluators.

One area of future work involves the extraction of code-specific features. One simple way to begin this process would be to learn from the coding manuals themselves. For example, the SSIRS states that “sighs” are a relevant cue for a spouse’s level of sadness (Appendix). Therefore, we could train a detector to automatically recognize instances of sighs from the audio signal, which could then act as one informative feature for predicting sadness. In addition to learning from the coding manuals, we also want to incorporate greater insight from expert psychologists into the computational modeling framework for each of the behavioral codes.

Other future plans include incorporating spouse and code correlations (see Table 3.2) in the modeling framework by jointly predicting the codes, rather than treating each independently. One possible direction is to use graphical models (e.g., Bayesian networks) that directly model inter-code and spouse dependencies. Another option is to develop a two-stage classification scheme; the first stage would classify each code independently, and the second stage would combine the output hypotheses to exploit the code and spouse correlations.

While we performed one experiment that analyzed the more “ambiguous” instances that fell in the middle 60% of the code range, the focus of this chapter was on classifying the extreme instances. Our future work will move away from this binary classification problem and concentrate on modeling and predicting *all* the instances. Toward this goal, we will experiment with regression techniques (that treat the code scores in a continuous manner) and ordinal regression techniques, which treat the code scores in an ordinal manner (e.g., [150]). This future work will also model and predict individual evaluator code scores, as opposed to using only the average scores across all evaluators.

While the availability of transcriptions enabled us to employ speech-text alignment to segment the corpus by speaker, we also plan to experiment with fully automatic ways to pre-process the Couple Therapy corpus. State-of-the-art automatic speaker diarization algorithms will be used to segment the audio into speaker-specific regions (e.g., [92,167]). In addition, source separation techniques may prove useful in detecting regions of overlapped speech, which may be another relevant cue/feature for predicting the behavioral codes.

We also hope to adopt a more multimodal approach to predicting the behavioral codes. As seen in the Appendix, the evaluators are trained to look for a variety of visual gestural cues (e.g., eye gaze, head orientation, facial expressions such as smiling and scowling, bodily expressions such as arms crossing). Thus, it is important to sense, model, and analyze relevant video information if we are to accurately code behavioral data. While the Couple Therapy corpus may not be ideal for this research due to the low data quality of the videos (see Section 3.2), we are in the process of collecting multimodal data of dyadic discussions in a "smart room" outfitted with multiple high-quality audio-video sensors [151].

The results of the current study open new avenues for exploration in couples research as well as new possibilities for intervention that would not otherwise be possible. For example, co-author Christensen is a member of a research effort that is evaluating the efficacy of IBCT delivered via the web. A primary aim of the project is to make IBCT broadly available to couples who may otherwise have difficulty or be hesitant about seeking couple therapy. Our goal is to apply and extend the findings and methods of the current study to enable couples receiving IBCT over the web to get automated feedback about their own behavior by submitting a recorded sample of behavior over the web.

We are also considering extending the methods and findings of the current study to providing near real-time feedback and intervention to couples who engage in moderate levels of intimate partner aggression. Though conflict is one of the most well replicated predictors of intimate partner aggression, couples frequently have difficulty recognizing when they are exhibiting conflict-instigating behaviors that increase risk for aggression. The methods and findings of the current study could be used to provide feedback to aggressive couples using smart phones or other mobile devices that allow for audio sampling.

We are also working with psychologists on a number of other BSP application domains: autism, depression, addiction, and post traumatic stress disorder. Collaborating at an earlier stage in the research has enabled us to develop hypotheses and design experiments that benefit both psychology and engineering. We hope that these ongoing and future BSP endeavors will promote synergistic collaborations between engineers and psychologists and ultimately push both fields forward.

3.10 Appendix: Coding manual written guidelines

Below we provide the written guidelines of the six codes analyzed in this dissertation, copied from the two coding manuals. “Acceptance of other” and “blame” were from the Couples Interaction Rating System (CIRS) [93], and “global positive,” “global negative,” “sadness,” and “use of humor” were from the Social Support Interaction Rating System (SSIRS) [99].

Acceptance of Other Indicates understanding and acceptance of partner’s views, feelings, and behaviors. Listens to partner with an open mind and positive attitude. May paraphrase partner’s statements. Subject need not agree with the partner’s views, but respects these views. Anger and criticism imply low acceptance, but high acceptance (scores of 8,9) goes beyond a lack of criticism and includes warmth toward partner. Resignation (i.e., settling unenthusiastically for a situation that you don’t believe will change) should not be considered acceptance.

Blame Blames, accuses, or criticizes the partner, uses critical sarcasm; makes character assassinations such as, “you’re a real jackass,” “all you do is eat,” or “why are you such a jerk about it?” Explicit blaming statements (e.g., “you made me do it,” or “you prevent me from doing it”), in which the spouse is the causal agent for the problem or the subject’s reactions, warrant a high score.

Global Positive An overall rating of the positive affect the target spouse showed during the interaction. Examples of positive behavior include overt expressions of warmth, support, acceptance, affection, positive negotiation, and compromise. Positivity can also be expressed through facial and bodily expressions, such as smiling and looking happy, talking easily, looking comfortable and relaxed, and showing interest in the conversation.

Global Negative An overall rating of the negative affect the target spouse shows during the interaction. Examples of negative behavior include overt expressions of rejection, defensiveness, blaming, and anger. It can also include facial and bodily expressions of negativity such as scowling, crying, crossing arms, turning away from the spouse, or showing a lack of interest in the conversation. Also factor in degree of negativity based on severity (e.g., a higher score for contempt than apathy).

Sadness Expression of sorrow and grief or resignation. Sadness is most apparent from behavioral cues, such as tearing or crying, looking down and dejected, sighing, speaking in a soft or low tone, and holding the head down. Verbalizations can involve expressing low spirits, unhappiness, and disappointment.

Use of Humor Measures the use of positive, non-derisive humor to lighten the mood during the interaction for both the target and non-target spouse. This can include jokingly making fun of the self, lightly teasing the spouse, or making a reference to a mutually shared joke. This category would not include making a joke at the expense of the self or spouse, mocking, or being sarcastic. If the target spouse does not initiate the humor but reacts positively to the other spouse's humor, code a low score.

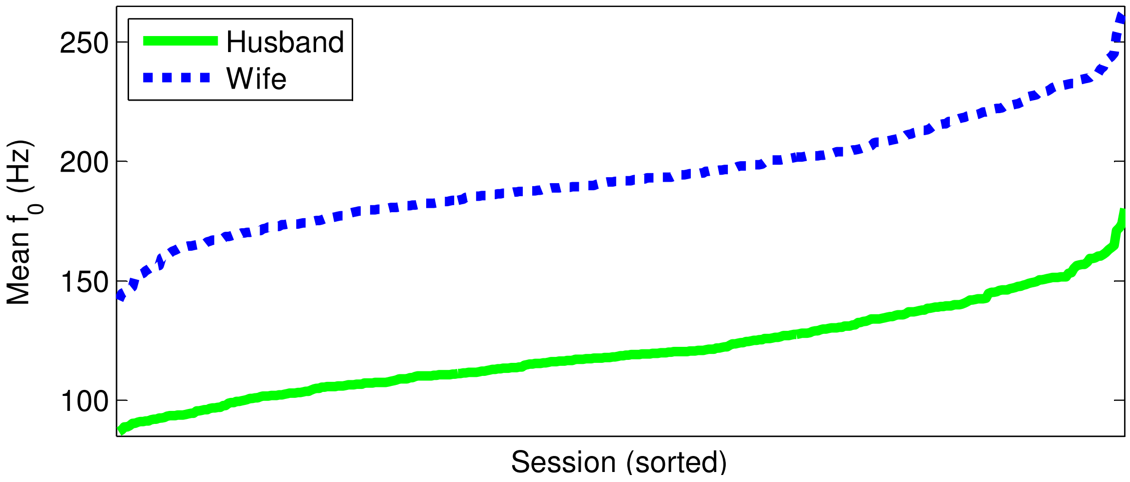

Figure 3.6: The ordered mean fundamental frequency (f_0) estimates for the wife and husband in each of the 372 coded sessions that met the SNR and speaker segmentation criteria.

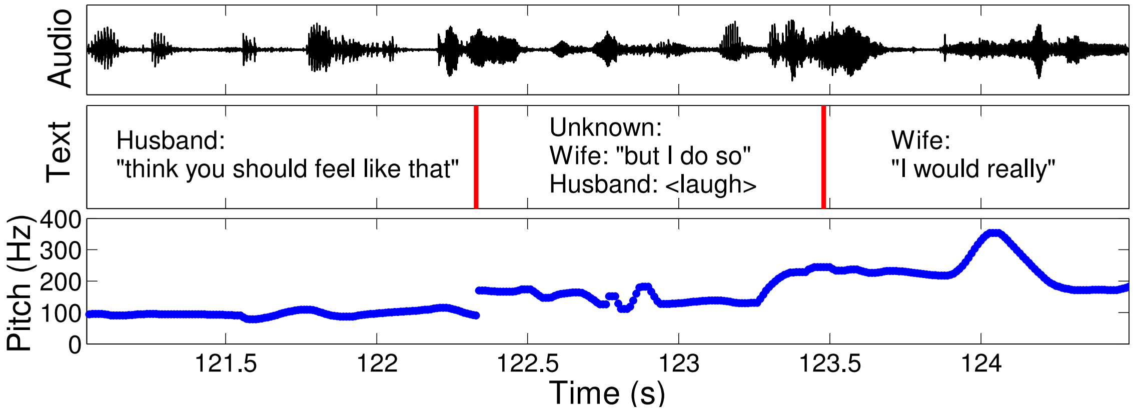

Figure 3.7: Example of the speaker segmentation and processed f_0 signal. In this particular example, the middle portion (labeled "Unknown") was unable to be automatically segmented due to overlapped speech (the husband was laughing while the wife was speaking).

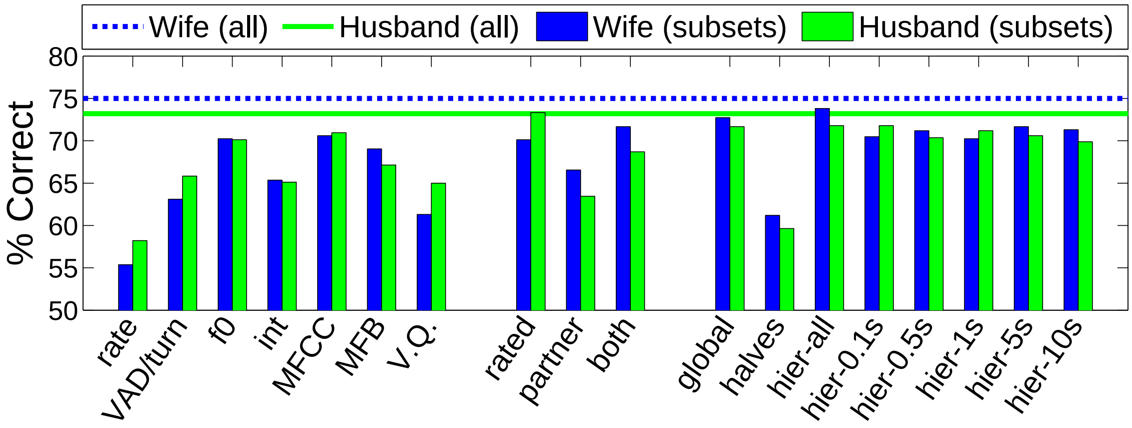

Figure 3.8: Average percentage of correctly classified instances across the six codes for the wife and husband (using gender-specific models and l^2-regularized logistic regression) for single feature *subsets* and compared to the case when we used *all* the features.

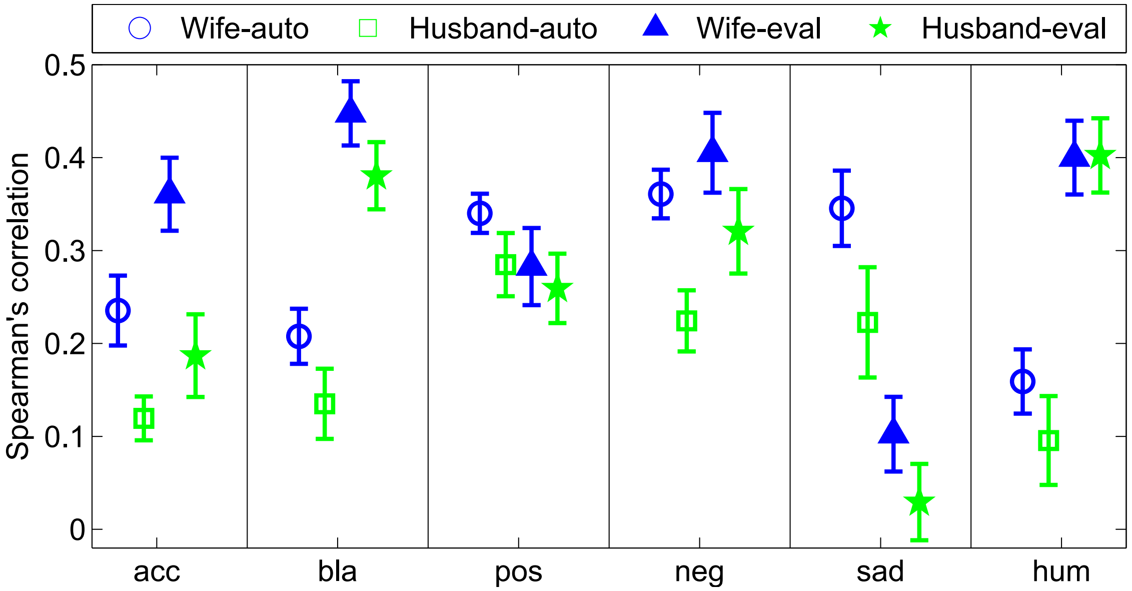

Figure 3.9: Performance in ranking the instances of the middle 60% of the six behavioral codes for the automatic (auto) system, as compared to inter-evaluator agreement (eval). Each plot shows the mean and standard deviation in the Spearman's correlation.

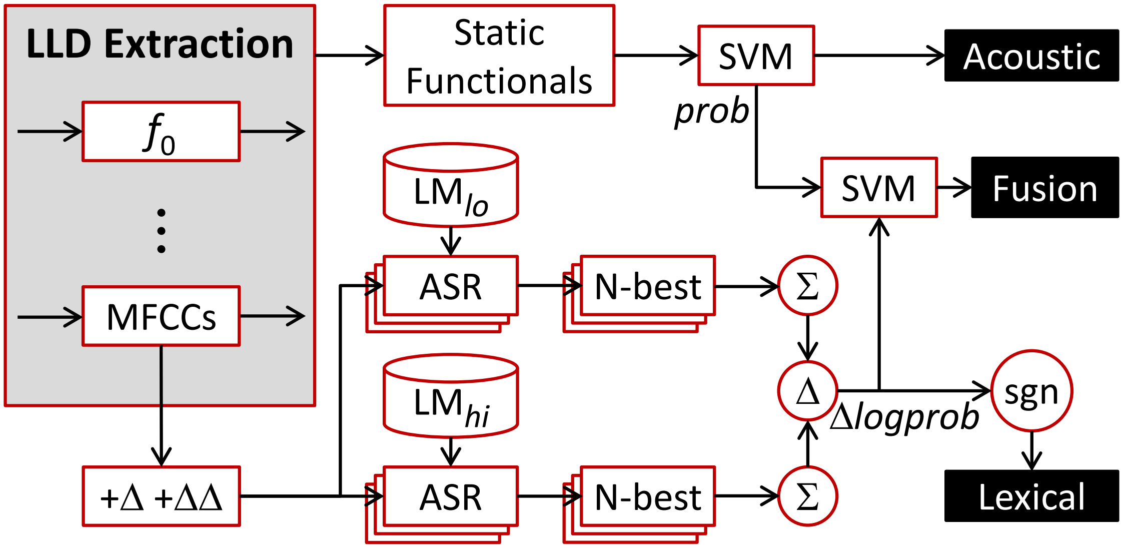

Figure 3.10: System block diagram, from the low-level descriptors (LLDs) to the blame class outputs for the static acoustic classifier, ASR-derived lexical classifier, and fusion classifier.

Chapter 4

Autism Diagnosis

4.1 Introduction

[1] Autism spectrum disorders (ASD) are highly heritable neurodevelopmental disorders characterized by a triad of core deficits, including impaired social behaviors, communication, and restricted/repetitive behaviors [3]. ASD is considered a "spectrum" disorder because symptomatology severity in each of the core domains can vary greatly. There have been increased research efforts in ASD, as recent prevalence studies indicate that as many as 1 in 110 children are diagnosed with ASD [138]. Studies have shown that early diagnosis and intensive early intervention can lead to improved social and communication skills in autistic children [54].

Psychologists in both research and in practice rely heavily upon observational methods for the assessment of social and communicative abilities. The Autism Diagnostic Observation Schedule (ADOS) is one of the most widely used clinical research instruments for the assessment and diagnosis of ASD and is appropriate for individuals with

[1] This research was supported in part by the National Institute of Child Health and Human Development, the National Science Foundation, Autism Speaks, and the Marino Autism Research Institute.

varying ages and verbal abilities [82, 118]. The semi-structured 30-60 minute interaction provides a trained psychologist with behavioral evidence that can be evaluated along dimensions important in diagnosing autism. Because of the qualitative descriptions of the assessments and the lack of continuous quantitative measures, one challenge in using the ADOS (and with observational methods in general) is the subjective nature inherent to the rating system.

Technology can assist with this process in a number of ways. Audio-video sensors can record the child-clinician interaction in a consistent fashion, and state-of-the-art signal processing methods can facilitate quantitative analyses and modeling using the audio-video data. Computational methods may be better suited than human observers in quantitatively tracking certain human behavioral cues (e.g., speech prosody, hand gestures). In recent years, there have been new emerging fields (e.g., social signal processing [174], behavioral signal processing [18]) concentrating on robustly measuring high-level human behaviors during realistic interactions using audio-video data. These data-driven signal cues could provide researchers and clinicians with a quantitatively dynamic source of information. We wish to emphasize that the clinical acumen of experienced psychologists is invaluable in evaluating and diagnosing children with ASD, and that we anticipate our work with this corpus will augment, rather than supplant, an expert clinician.

The collection of realistic corpora is a critical step in many data-driven engineering pattern analysis and recognition realms. Example domains include automatic speech recognition [80, 159], affect/emotion recognition [34], and automatic literacy assessment [106]. In recent years, there has been significant engineering-related work on analyzing the speech and language of children with ASD; the experiments have occurred in a variety

Module	*Subject*	*Interaction subtasks*
1	< phrase speech	free play, response to name, response to joint attention, bubble play, anticipation of routine with objects, responsive social smile, anticipation of social routine, functional and symbolic imitation, birthday party, snack
2	phrase speech	construction task, response to name, make-believe play, joint interactive play, conversation, response to joint attention, demonstration task, description of picture, telling story from book, free play, birthday party, snack, anticipation of routine with objects, bubble play
3	fluent children	construction task, make-believe play, joint interactive play, demonstration task, description of picture, telling story from book, cartoons, emotions, social difficulties/annoyance, friends and marriage, loneliness, creating a story
4	fluent teens/adults	construction task, description of picture, telling story from book, cartoons, emotions, social difficulties/annoyance, friends and marriage, loneliness, creating a story, daily living, current work/school, plans and hopes

Table 4.1: A list of the interaction subtasks for each ADOS module [118], along with the intended language level and/or age of the subject.

of social contexts, ranging from isolated speaking tasks [171, 172] and structured clinical assessments [94] to unconstrained home environments [142]. However, since ASD affects vocal, linguistic, and gestural social behavioral patterns, there is a need for multimodal data of children with ASD.

Towards this end, we introduce the *USC CARE Corpus* [26], comprised of real, spontaneous child-psychologist interactions, recorded in a controlled clinical environment in the context of the ADOS. The collection of this corpus is the necessary first step in analyzing complex social interactions between expert psychologists and children with ASD. We plan to use the audio-video data for a number of multimodal signal processing research projects, ranging from improved modeling of children's spontaneous speech to the analysis of atypical communication patterns and the study of dialogs in clinical settings.

In addition to offering a new problem domain for engineering, this unique corpus has important potential contributions to the ASD community. Ultimately, this research could help support ASD diagnoses with quantifiable and adaptable metrics, provide more accurate stratification of subgroups for targeted interventions, and automatically track children's progress during the treatment. Section 4.2 discusses the ADOS interaction paradigm in more detail. We describe the USC CARE Corpus (an ongoing data collection) in Section 4.3, and we explain our initial analyses and future intended work in Section 4.4. Related work and efforts are discussed in Section 4.5, and we conclude in Section 4.6.

4.2 ADOS Interaction Paradigm

The ADOS is a "gold-standard" research tool for the assessment of the triad of behaviors that together are diagnostic for ASD [82, 118]. There are four modules; the psychologist determines which one to administer, depending on the subject's expressive language level and chronological age (see Table 4.1).

During the ADOS, the subject interacts spontaneously with a psychologist (and a parent for modules 1 and 2) for approximately 30 to 60 minutes. To ensure that the interaction is standardized, the psychologist follows a predetermined semi-structured set of subtasks. Table 4.1 lists the interaction subtasks for each ADOS module. This table shows that there is significant subtask overlap between the four modules, with more conversational and interview-style subtasks for the fluent-speaking subjects in modules 3 and 4. Modules 3 and 4 are typically administered at a table, while modules 1 and 2 require the child (and parent) to move around the room.

The ADOS was designed for psychologists to make assessments on the subject's proficiency for a number of communication and social interaction skills. Observations are noted by the psychologist in real-time during the interaction, and the psychologist rates the subject's behavior immediately after the session according to the module-specific ADOS coding manual. Each coding manual consists of approximately 28 codes, which are broken down into five main groupings (e.g., communication, reciprocal social interaction, play/imagination/creativity, stereotyped behaviors/restricted interests, and other abnormal behaviors). The codes assess speech (e.g., speech abnormalities/atypical prosody), language (e.g., stereotyped/idiosyncratic use of words/phrases), nonverbal communication (e.g., directed facial expressions, eye contact, use of gestures), and other behaviors (e.g., imagination/creativity, overall quality of rapport).

Each code has a written description, and the psychologist chooses the value that best describes the subject's behavior; while in some cases codes are based on single subtasks, most codes consider overall behavior throughout the evaluation. The summary ADOS algorithm includes those codes that were shown in the standardization research to best predict an autism diagnosis. The algorithm codes are summed to attain communication and social interaction subtotals (each with predetermined autism and ASD cut-offs), and a total score is computed for a final ADOS classification. In addition, the clinician administering the test is asked to give an overall diagnosis of autism/ASD, based on the ADOS scores as well as other information that may influence the validity or interpretation of the scores [118].

Psychologists are trained to administer and code the ADOS using a stringent training protocol that includes reaching agreement on multiple ADOS protocols with an expert evaluator. This training process is time-consuming and challenging. One of the main challenges is due to the qualitative nature for some of the codes in the ADOS coding manuals. For example, in modules 2-4, speech abnormalities (i.e., atypical prosody) are coded based on descriptions such as, “Little variation in pitch and tone; rather flat or exaggerated intonation, but not obviously peculiar, or slightly unusual volume, and/or speech that tends to be somewhat unusually slow, fast, or jerky.” From this qualitative description, it is clear that coding for prosodic abnormalities requires knowledge of normative prosody and the range of atypicalities associated with ASD. Thus, it takes a great deal of specialized training to learn to reliably score the ADOS. Moreover, the scoring is categorical and does not provide continuous measures that can be utilized for population stratification.

Our goal is to contribute to overcoming some of these limitations through the development of engineering algorithms and tools that are based on the analysis of a large corpus of ADOS sessions. Incorporating engineering methods that quantitatively assess communication/social interaction behaviors could help support scalability and the analysis and decision capabilities of psychologists. In addition, it has the potential to contribute to research aimed at better understanding variations in the communication and social patterns of children with ASD.

4.3 The USC CARE Corpus

4.3.1 Background

The Center for Autism Research in Engineering (CARE) was established in 2009 at the University of Southern California (USC), with the goal to better incorporate engineering and computer science methodologies into autism research through interdisciplinary collaborations. Early and ongoing work between the Signal Analysis and Interpretation

Laboratory (SAIL) at USC and the USC University Center for Excellence in Developmental Disabilities at Children's Hospital Los Angeles (CHLA) experimented with child-computer interaction applications [15, 134].

More recently, we have teamed up with researchers at the Zilkha Neurogenetic Institute and the Boone Fetter Clinic (BFC) at CHLA to record ADOS evaluations for an ongoing prospective clinical and genetics study on the relationship between ASD and gastrointestinal dysfunction (GID). The subsequent ADOS sessions from the Los Angeles-based families who agreed to be recorded for the study make up the *USC CARE Corpus*.

4.3.2 Recruited Subjects

All recruited subjects with a prior clinical diagnosis of ASD were administered the ADOS (to verify the diagnosis). The subjects were required to be between 5 and 18 years of age to participate in the study. In addition, all participating families were required to be native speakers of either English or Spanish. As part of the study, the parents filled out a number of standardized questionnaires (e.g., on their child's verbal abilities and social functioning), which are included as part of the corpus.

We began recording the ADOS evaluations in April, 2010. As of March, 2011, we have collected data from 70 subjects; our goal is to record 100 subjects. Table 4.2 provides statistics on the participants whose demographics have been uploaded to the database. The majority of the recruited families were native speakers of Spanish because the experiments took place in Los Angeles, California. Note that the gender imbalance in recruited subjects is due to the fact that males are four to five times more likely than females to be diagnosed with ASD [138].

Category	*Count/Statistic*
Age (years)	*mean*: 9.3, *std. dev.*: 3.1, *range*: 5.2-17.0
Gender	*male*: 49, *female*: 11
Native language	*Spanish*: 38, *English*: 22
Ethnicity	*Hispanic/Latino*: 34, *White*: 9, *Other*: 8, *unk*: 9
ADOS module	*#1*: 17, *#2*: 13, *#3*: 28, *#4*: 2
ADOS diagnosis	*autism*: 37, *ASD*: 9, *no ASD*: 8, *unk*: 6

Table 4.2: Demographic statistics of the 70 recorded subjects administered the ADOS (as of March, 2011). The unknown ("unk") entries have not been uploaded to the database yet.

4.3.3 ADOS Codes

Three research-certified psychologists administered the ADOS evaluations; co-author M. E. Williams was the lead psychologist, and she oversaw the training of the other two psychologists. The administering psychologist coded the subject according to the module-specific ADOS manual (Section 4.2). The resulting code scores are a critical part of the USC CARE Corpus, since they represent standardized expert coding of ASD-relevant behaviors. We also have the final ADOS diagnosis from the psychologist; most of the subjects met the autism or ASD cut-offs (Table 4.2).

Some of our intended future work will make use of these code scores and final ADOS classification. This information can be used to cluster children with similar characteristics and/or to facilitate the use of supervised learning techniques to automatically categorize typical from atypical behavior. As part of our future work, we may also collect similar ADOS data from typically developing children, which would allow us to train normative behavioral models.

4.3.4 Multimodal Data Collection

All ADOS evaluations took place in the BFC at CHLA, with floor dimensions of 3.3m x 2.3m and ceiling height of 2.6m. The BFC is a shared multi-use clinical space, so we used a portable smart-room solution with multiple audio-video sensors to unobtrusively record the interaction. All of the sensors operated in the far-field to ensure that they were not disruptive to the natural flow of the interaction and to maximize the ecological validity of the experiments. In addition, we did not place any sensors directly on the subject or the subject's clothing, to minimize the possibility of exacerbating anxiety states (many individuals with ASD have anxiety and sensory sensitivities [138]).

Two Sony HDR-SR12 HD Handycam Camcorders were mounted on tripods approximately 1m off the floor and 2m from the subject. They operated in the corners of the room opposite to the child to capture the child's body and face while seated at the table. See Figure 4.1 for the layout of the clinical room. We used the highest quality video settings: 1080i resolution with a 16:9 widescreen aspect ratio (1920x1080 pixels, 59.94 interlaced frames per second, H.264/MPEG-4 AVC compression).

We recorded stereo audio signals from each camcorder's internal microphones (48 kHz, 16-bit). In addition, we recorded audio from two high-quality directional shotgun microphones (SCHOEPS CMIT 5 U), which were mounted next to the camcorders. We used the Edirol R-4 Pro recorder to capture the uncompressed audio (48 kHz, 24-bit).

The audio-video equipment takes approximately ten minutes to set up and five minutes to dismantle. All audio-video signals were synchronized to the nearest frame of video by clapping before and after the sessions and manually marking these clap times in each of the audio-video signals. We currently have 50 hours of data (for each of the two channels of video and six channels of audio); the average session duration is 49.5 minutes.

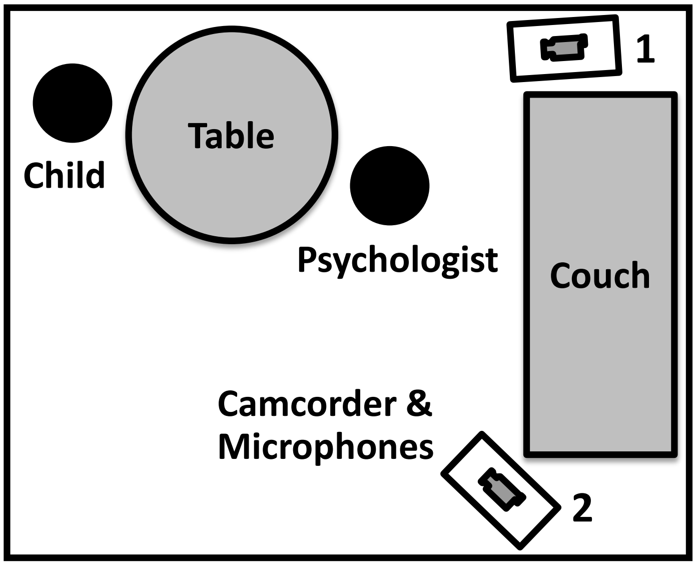

Figure 4.1: The layout of the clinical room, showing the location of the participants and the placement of the camcorders and microphones.

4.4 Initial Analyses & Future Work

4.4.1 Manual Transcription & Language Processing

We are in the process of manually transcribing and segmenting the USC CARE Corpus into individual speaker turns; these transcribers are blinded to the code scores and final ADOS classification. The resulting transcriptions will provide us reference landmarks indicating when each person was speaking, the word-level lexical content of the speech, and enriched transcription of: partial-words, stuttering, disfluencies, nonverbal *communication* (e.g., laughs, sighs), nonverbal *vocalizations* (e.g., grunts, babbles), mispronunciations, and neologisms. In addition, each utterance is labeled as a question, fragment, interruption, and/or complete thought. The transcription manual we devised was adapted from the SALT transcription guidelines [127], used in many ASD-related studies.

We plan to use these transcriptions to analyze the children's language use, turn-taking trends, and other surface behaviors, as was done in [94]. Importantly, the transcriptions can also be used to train a number of speech-related signal processing tools (voice activity detector, acoustic models, language models) for the specific environmental conditions and speaker demographics of the corpus.

4.4.2 Speech Signal Processing

Speech is one modality in which state-of-the-art signal processing can make a profound impact. As highlighted in Section 4.2, the assessment of children's speech is one important aspect of the ADOS. There are a variety of atypicalities associated with the prosody of verbal autistic children. Studies have suggested that individuals with ASD have problems with lexical stress and pragmatic prosody [121]. Others have reported durational abnormalities, with the speech either too fast or slow [59]. Consistent with Kanner's description more than 60 years ago [102], current listeners often report a "bizarre" quality to the speech (e.g., monotonous intonation) [74, 121].

One of our future goals with this corpus is to develop automated methods to assess the various dimensions of prosody (e.g., rate, intonation, volume) within the context of the ADOS. While significant work has been done on assessment of prosody during children's constrained speaking tasks [171, 172], the USC CARE Corpus presents new challenges and opportunities due to the spontaneous nature of the speech. As part of this future work, we will also have to account for the rich linguistic diversity of the recruited subjects (Table 4.2).

We also plan to develop methods to automatically detect nonverbal vocalizations (e.g., grunts, shrieks, babbles) during the ADOS. It was found that these non-lexical vocalizations were a key feature in automatically separating typically developing children from those diagnosed with ASD [142]. This is a challenging learning problem, since these behaviors are idiosyncratic and can be rarely occurring events.

4.4.3 Multimodal Signal Processing

One of the unique aspects of the USC CARE Corpus is the multiple channels of synchronized audio and video. As discussed in Section 4.2, the assessments for the ADOS, and the social communication impairments for ASD, are inherently multimodal. In fact, some of the behavioral abnormalities are due to atypical synchrony between expressive modalities. Therefore, multimodal signal processing techniques are needed to fully capture the unique differences in communication of children with ASD.

Our initial plans will concentrate on processing specific subtasks within the ADOS interaction (Table 4.1). We will first consult with the psychologists to discover the cues most relevant for the subtasks. This will help inform the multimodal feature extraction and an appropriate automatic learning method. We can validate our methods using the psychologists' codes (Section 4.3.3). In some cases, we will manually code events at a finer-grained temporal scale, which we can use to train (and evaluate) specific behavioral detectors/classifiers. Combining data-driven and expert-inspired knowledge in a multimodal signal processing framework also is an area of future work.

4.5 Related Work

There has been some work in recent years on the use of computational methods to study autistic children's speaking patterns and prosody. For example, van Santen and his colleagues have worked on automatic assessment of affective prosody through a comparative study of within-speaker productions of minimal pairs. A minimal pair is a word like, "present," where there are two different lexical stress syllables (pre'sent or 'present), but the underlying phonetic transcription of the two are identical. In this research, children diagnosed with ASD and typically developing children were recording speaking a number of minimal pairs by imitating pronunciations or by correcting an incorrectly read minimal pair (based on word context). It was shown that by using prosodic features and a dynamic time warping method, the prosodic differences between these minimal pair pronunciations could be quantified. Their studies revealed that children with ASD did, in fact, pronounce the minimal pairs differently, but that their pronunciations differed from those pronounced by the typically developing children. Specifically, it was shown that the two populations of children differed in their balance between the various prosodic cues [171, 172]. One limitation of these studies is the non-spontaneous elicitation of the prosodic patterns. This may have caused the speech samples to be less representative of the real manner in which the children communicate.

There has also been significant work on speech analysis of autistic children from the LENA Foundation, where researchers have taken a more unsupervised approach to the problem [142]. In this work, children are outfitted with a microphone they wear in their clothes, which records their voice and all other sounds (including background noises and the speech of other people) in the child's natural home environment. Each child is recorded for a period of weeks to months. The child's speech is automatically segmented from the rest of the audio recording, resulting in a large-scale sample of the child's speech. A 12-dimensional feature set, representing various high-level characteristics of the children's voice (e.g., frequency of "growl"-like vocalizations) are then used to separate children diagnosed with autism to ones with language delay only (and a negative diagnosis of autism) and ones with neither (which were considered typically-developing in this study). While this research holds great promise toward the ability to collect affordable and natural speech over a long period of time, it does not benefit from the constrained interactions that occur in a clinic. Psychological methods have been devised to constrain a situation to extract relevant information about the child's social communication, and these situations are not applied in an unsupervised scenario like a child's home.

4.6 Conclusions

In this chapter, we introduced the USC CARE Corpus, a large multimodal corpus of ADOS evaluations. This data is important to facilitate the analysis of complex interactions involving children with ASD while in a controlled clinical environment. In addition to describing the unique elements of the ADOS and the multimodal recording set-up, we also provided an outline for future work in speech, language, and multimodal signal processing with this novel corpus.

We believe that the collection of the USC CARE Corpus represents a key step towards better incorporating engineering methodologies into the behavioral sciences and healthcare related domains, including neurodevelopmental disorders. This is a primary goal of behavioral signal processing.

The BSP framework fits nicely with this problem, since it can help quantify these difficult-to-describe symptoms of autism using objective signal-based cues within the context of the ADOS. Psychologists can make use of these tools to inform their decisions regarding the diagnosis and treatment of social disorders. Our specific goal is to automate aspects of the ADOS coding (e.g., atypical prosody) using data-driven methods trained on the novel corpus we are currently collecting. Incorporating quantitative methods and models could lead to a more consistent grading scheme across subjects and over time. The technology could potentially be scalable to large populations of children.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This thesis describes completed research, centered around the quantification and prediction of human subjective judgments on various aspects of human behavior. Specifically, we examined this research problem in the context of three real-life applications: literacy assessment, couple therapy research, and autism diagnosis.

Automatic literacy assessment technology can help children acquire reading skills by providing teachers valuable feedback in a repeatable, consistent manner. There is a need for more high-level automatic assessments that capture the overall performance of the children. These high-level assessments can be viewed as an interpretive extension to lower-level assessments, and may be more perceptually relevant to teachers and helpful in tracking performance over time. In this thesis, we modeled and predicted the overall reading ability of young children reading a list of English words aloud. This research was broken into two main parts. The first part was a user study, in which 11 human evaluators rated the children on their overall reading ability based on the audio recordings.

In the second part, we ran machine learning experiments to predict evaluators' scores using features automatically extracted from the audio. The features were human-inspired and correlated with cues human evaluators stated they used: pronunciation correctness, speaking rate, and fluency. We investigated various automated methods to verify the correctness of the word pronunciations and to detect disfluencies in the children's speech using held-out annotated data. Using linear regression techniques and by incorporating evaluator variability, we automatically predicted individual evaluators' high-level scores with a mean Pearson correlation coefficient of 0.828, and we predicted average evaluator's scores with correlation 0.952. Both these human-machine agreement statistics exceeded the mean inter-evaluator agreement statistics.

One of the goals of behavioral signal processing is the automatic prediction of relevant high-level human behaviors from complex, realistic interactions. Observational methods are fundamental to the study of human behavior in the behavioral sciences. For example, in the context of research on intimate relationships, psychologists' hypotheses are often empirically tested by video recording interactions of couples and manually coding relevant behaviors using standardized coding systems. This coding process can be time-consuming, and the resulting coded data may have a high degree of variability because of a number of factors (e.g., inter-evaluator differences). These challenges provide an opportunity to employ engineering methods to aid in automatically coding human behavioral data. In the second case study of this thesis, we analyzed a large corpus of married couples' problem-solving interactions. Each spouse was manually coded with multiple session-level behavioral observations (e.g., level of blame toward other spouse), and we used acoustic speech features to automatically classify extreme instances for six selected codes (e.g., “low” vs. “high” blame). Specifically, we extracted prosodic, spectral, and voice quality features to capture global acoustic properties for each spouse and trained gender-specific and gender-independent classifiers. The best overall automatic system correctly classified 74.1% of the instances across the six selected codes. In addition, since many important behaviors can be conveyed through various communicative channels (e.g., speech, language, gestures), we compared two different classification methods with the “blame” code: the first classifier was trained with the conventional static acoustic features and modeled “how” the spouses spoke, and the second was a novel automatic speech recognition-derived classifier, which modeled “what” the spouses said. We got the best classification performance on the “blame” code (82% accuracy) by exploiting the complementarity of these acoustic and lexical information sources through score-level fusion of the two classification methods.

Finally, in the third case study examined in this thesis, we introduced the USC CARE Corpus, comprised of spontaneous and standardized child-psychologist interactions of children with a diagnosis of an autism spectrum disorder (ASD). The audio-video data was collected in the context of the Autism Diagnostic Observation Schedule (ADOS), which is a tool used by psychologists for a research-level diagnosis of ASD for children. The interaction consists of developmentally appropriate semi-structured social activities, providing the psychologist with a sample of behavior used to rate the child on a series of autism-relevant symptoms. Our future goal with this multimodal corpus is to investigate how analytical technology (e.g., speech and language processing) can enhance this observational rating task and provide greater insight into social behavior and communication. In Chapter 4, we provided demographic statistics on the recruited children (70 to date), described the multimodal recording set-up, and discussed current and future work for this novel corpus.

This dissertation contributes to a large framework, behavioral signal processing, which attempts to understand human behavior by modeling both the internal state of a person and observational processes and help support human experts' decision capabilities with new quantitative tools and measures.

5.2 Open Problems and Future Work

There are numerous technological challenges, as outlined in Chapter 1, with automatically quantifying and predicting subjective judgments made on human behavior. In addition, there are a number of open problems in the emerging area of behavioral signal processing. These challenges and open problems include the modeling of multiple sources of variability: heterogeneity in the displays of human behavior ("production") and subjectivity in the judgments of human behavior ("perception"). In addition, there is information across multiple modalities and cues, which can be distributed across various signals (e.g., acoustics, language, gestures), and it is not always clear how humans leverage this information to ultimately make holistic judgments. Finally, there is always the issue of realistic noisy data. This provides many future research opportunities in the development of robust signal processing and machine learning methodologies to model these complex subjective observational processes.

We tackled many of these challenges in this thesis through the analysis of specific problem domains in education, family studies, and health. We discuss future work for each of the three application domains at the ends of Chapter 2 (automatic literacy assessment), Chapter 3 (couples therapy research), and Chapter 4 (autism diagnosis). If we can continue to successfully address and overcome some of these challenges and open problems, this automatic framework has many tantalizing possibilities for society. For example, automatic literacy assessment tutors could revolutionize the classroom dynamic; smart-room sensing could interpret the communication patterns of multi-person interactions to help psychologists test hypotheses that would otherwise be infeasible to test; and objective, signal-based behavioral markers for developmental disorders like autism could be devised.

Moving forward, it is vital to continue to collect and analyze naturally-occurring human behaviors from multimodal data recorded in ecologically valid settings. One focus should be on developing computational methods that learn from experts in other fields; this is a challenging problem, since it is non-trivial to define optimal ways to transfer knowledge from experts (human observers with potentially years of training and experience) to machines. Humans and computers have different skill sets and can offer complementary information, so in some cases, it may be essential to develop technologies that collaboratively assist humans in enhancing their analysis capability and capacity. Ultimately, these technologies should be embraced by people and be considered useful, so incorporating people in the design process becomes critical. Future human-in-the-loop experiments can study ways in which computers can iteratively learn from experts and pinpoint where difficulties and limitations arise. Hopefully these models can exploit people's and machine's mutual strengths in processing behavioral data.

In addition to creating computational solutions in realistic experimental settings, future research efforts must also be centered on the critical step of incorporating these automated methods for use in real-life, societally-significant applications. One potential application of this work is providing psychologists and clinicians with near real-time feedback of important behavioral analytics of their subjects/patients. Having objective, quantitative cues at their disposal could transform the way human behavior is analyzed and provide psychologists/clinicians with a set of tools that was not possible before (or was possible only after processing the data manually). Another application is in enriching human-machine interactions, which are becoming more prevalent due to the increased usage of computers and portable electronic devices. Enabling computers to detect and appropriately respond to important human behaviors (e.g., confusion, frustration, uncertainty) could improve the naturalness and effectiveness of human-machine interactions. BSP solutions could be employed for widespread use in various technologies (e.g., user interfaces, interactive voice response systems), providing an invaluable link for human interlocutors.

The emerging field of behavioral signal processing (BSP) promises to be an important step in creating new possibilities for human-centered engineering. It is my hope that this dissertation provides a solid automatic framework and computational foundation, upon which future research can build.

References

[1] J. K. Aggarwal and S. Park, “Human motion: Modeling and recognition of actions and interactions,” in *2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings*, 2004, pp. 640–647.

[2] A. Alwan, Y. Bai, M. P. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, “A system for technology based assessment of language and literacy in young children: The role of multiple information sources,” *Proceedings of MMSP, Greece*, 2007.

[3] American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, “Washington, dc, american psychiatric association,” vol. Fourth ed., 2000.

[4] D. C. Atkins, S. A. Milbright, A. Dueck, K. Reimer, and A. Christensen, “The language of therapy: The promises and hurdles of computational linguistics,” in *Annual Meeting of the Association for Behavioral and Cognitive Therapies*, Washington, D.C., Nov. 2005.

[5] K. Audhkhasi and S. S. Narayanan, “Data-dependent evaluator modeling and its application to emotional valence classification from speech,” in *Proc. of Interspeech*, 2010.

[6] ——, “Emotion classification from speech using evaluator reliability-weighted combination of ranked lists,” in *Proc. of ICASSP*, 2011.

[7] A. Batliner, M. Blomberg, S. D’Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, “The PF_STAR Children’s Speech Corpus,” in *Ninth European Conference on Speech Communication and Technology.* Citeseer, 2005.

[8] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, *The Automatic Recognition of Emotions in Speech*, ser. Emotion-Oriented Systems: The Humaine Handbook Cognitive Technologies, 2011, pp. 71–99.

[9] B. R. Baucom, K. Eldridge, J. Jones, M. Sevier, M. Clements, H. Markman, S. Stanley, S. L. Sayers, T. Sher, and A. Christensen, “Relative contributions of relationship distress and depression to communication patterns in couples,” *Journal of Social and Clinical Psychology*, vol. 26, no. 6, pp. 689–707, 2007.

[10] B. Baucom, D. Atkins, L. Simpson, and A. Christensen, "Prediction of response to treatment in a randomized clinical trial of couple therapy: A 2-year follow-up," *Journal of Consulting and Clinical Psychology*, vol. 77, no. 1, pp. 160–173, 2009.

[11] D. H. Baucom, V. Shoham, K. T. Mueser, A. D. Daiuto, and T. R. Stickle, "Empirically supported couple and family interventions for marital distress and adult mental health problems," *Journal of Consulting and Clinical Psychology*, vol. 66, no. 1, pp. 53–88, 1998.

[12] J. G. Beck, J. Davila, S. Farrow, and D. M. Grant, "When the heat is on: Romantic partner responses influence distress in socially anxious women," *Behaviour Research and Therapy*, vol. 44, no. 5, pp. 737–748, 2006.

[13] M. P. Black, J. Chang, J. Chang, and S. Narayanan, "Comparison of child-human and child-computer interactions based on manual annotations," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction.* ACM, 2009, pp. 1–6.

[14] M. P. Black, J. Chang, and S. Narayanan, "An empirical analysis of user uncertainty in problem-solving child-machine interactions," in *Proceedings of the Workshop on Child, Computer and Interaction*, 2008.

[15] M. P. Black, E. Flores, E. Mower, S. S. Narayanan, and M. E. Williams, "Comparison of child-human and child-computer interactions for children with ASD," in *IMFAR*, 2010.

[16] M. P. Black, P. G. Georgiou, A. Katsamanis, B. R. Baucom, and S. S. Narayanan, ""You made me do it": Classification of blame in married couples' interactions by fusing automatically derived speech and language information," in *Proc. Interspeech*, Florence, Italy, 2011.

[17] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features," *Speech Communication*, 2011.

[18] M. P. Black, A. Katsamanis, C.-C. Lee, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *Proc. Interspeech*, 2010.

[19] M. P. Black, A. Kazemzadeh, J. Tepperman, and S. S. Narayanan, "Automatically assessing the ABCs: Verification of children's spoken letter-names and letter-sounds," *ACM Transactions on Speech and Language Processing*, vol. 7, no. 4, article 15, Aug. 2011.

[20] M. P. Black and S. S. Narayanan, "Improvements in predicting children's overall reading ability by modeling variability in evaluators' subjective judgments," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[21] M. P. Black, J. Tepperman, A. Kazemzadeh, S. Lee, and S. Narayanan, "Pronunciation verification of English letter-sounds in preliterate children," *in Proc. Interspeech*, 2008.

[22] ——, "Automatic pronunciation verification of English letter-names for early literacy assessment of preliterate children," *in Proc. ICASSP*, 2009.

[23] M. P. Black, J. Tepperman, S. Lee, and S. Narayanan, "Estimation of children's reading ability by fusion of automatic pronunciation verification and fluency detection," in *in Proc. Interspeech*, 2008.

[24] ——, "Predicting children's reading ability using evaluator-informed features," *Proc. Interspeech*, 2009.

[25] M. P. Black, J. Tepperman, S. Lee, P. Price, and S. Narayanan, "Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment," in *Proc. Interspeech*, 2007, pp. 206–209.

[26] M. P. Black, D. Bone, M. E. Williams, P. Gorrindo, P. Levitt, and S. S. Narayanan, "The USC CARE Corpus: Child-psychologist interactions of children with autism spectrum disorders," in *Proceedings of Interspeech, Florence, Italy*, Aug. 2011.

[27] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 1015–1028, 2011.

[28] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 7–74, 1998.

[29] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[30] M. B. Brown and A. B. Forsythe, "Robust tests for the equality of variances," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 364–367, 1974.

[31] M. Brüne, C. Sonntag, M. Abdel-Hamid, C. Lehmkämper, G. Juckel, and A. Troisi, "Nonverbal behavior during standardized interviews in patients with schizophrenia spectrum disorders," *The Journal of Nervous and Mental Disease*, vol. 196, no. 4, pp. 282–288, Apr. 2008.

[32] M. Bulut and S. S. Narayanan, "On the robustness of overall F0-only modifications to the perception of emotions in speech," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4547–4558, June 2008.

[33] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, and R. Huber, "Detecting real life anger," in *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, Taipei, Taiwan, 2009, pp. 4761–4764.

[34] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[35] C. Busso, Z. Deng, U. Neumann, and S. Narayanan, "Learning expressive human-like head motion sequences from speech," *Data-Driven 3D Facial Animation*, pp. 113–131, 2007.

[36] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, 2009.

[37] C. Busso, S. Lee, and S. S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, 2009.

[38] N. Campbell, "Databases of emotional speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[39] R. J. Carroll and D. Ruppert, *Transformation and Weighting in Regression.* New York, NY: Chapman & Hall, 1988, ch. 2: Generalized least squares and the analysis of heteroscedasticity.

[40] J. Cassell and K. Ryokai, "Making space for voice: Technologies to support children's fantasy and storytelling," *Personal and Ubiquitous Computing*, vol. 5, no. 3, pp. 169–190, 2001.

[41] C. C. Chang and C. J. Lin, *LIBSVM: A library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[42] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, Montreal, Quebec, Canada, May 2004, pp. 509–512.

[43] A. Christensen, D. C. Atkins, B. R. Baucom, and J. Yi, "Marital status and satisfaction five years following a randomized clinical trial comparing traditional versus integrative behavioral couple therapy," *Journal of Consulting and Clinical Psychology*, vol. 78, no. 2, pp. 225–235, 2010.

[44] A. Christensen, D. C. Atkins, J. Yi, D. H. Baucom, and W. H. George, "Couple and individual adjustment for 2 years following a randomized clinical trial comparing traditional versus integrative behavioral couple therapy," *Journal of Consulting and Clinical Psychology*, vol. 74, no. 6, pp. 1180–1191, 2006.

[45] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. H. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *J. of Consulting and Clinical Psychology*, vol. 72, pp. 176–191, 2004.

[46] A. Christensen and C. L. Heavey, “Gender differences in marital conflict: The demand/withdraw interaction pattern,” *Gender Issues in Contemporary Society*, vol. 6, pp. 113–141, 1990.

[47] A. Christensen, N. S. Jacobson, and J. C. Babcock, “Integrative behavioral couple therapy,” in *Clinical handbook of marital therapy*, 2nd ed., N. S. Jacobsen and A. S. Gurman, Eds. New York: Guilford Press, 1995, pp. 31–64.

[48] S. M. Chu and T. S. Huang, “Bimodal speech recognition using coupled hidden Markov models,” in *Sixth International Conference on Spoken Language Processing*, 2000.

[49] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, “Automatic pronunciation scoring of words and sentences independent from the non-native’s first language,” *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, 2009.

[50] P. Cosi, R. Delmonte, S. Biscetti, R. Cole, B. Pellom, and S. Vuren, “Italian literacy tutor-tools and technologies for individuals with cognitive disabilities,” in *InSTIL/ICALL Symposium 2004*, 2004, pp. 391–407.

[51] R. Cowie, “Perceiving emotion: Towards a realistic understanding of the task,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3515–3525, 2009.

[52] A. Coy and J. Barker, “An automatic speech recognition system based on the scene analysis account of auditory perception,” *Speech Communication*, vol. 49, no. 5, pp. 384–401, May 2007.

[53] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech,” *The Journal of the Acoustical Society of America*, vol. 111, p. 2862, 2002.

[54] G. Dawson, S. Rogers, J. Munson, M. Smith, J. Winter, J. Greenson, A. Donaldson, and J. Varley, “Randomized, controlled trial of an intervention for toddlers with autism: The Early Start Denver Model,” *Pediatrics*, vol. 125, no. 1, pp. 17–23, 2010.

[55] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *J. of the Acous. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[56] A. DeBruin-Parecki, “Evaluating early literacy skills and providing instruction in a meaningful context,” 2004.

[57] L. Devillers and N. Campbell, “Special issue of Computer Speech and Language on affective speech in real-life interactions,” *Computer Speech & Language*, vol. 25, no. 1, pp. 1–3, 2011.

[58] L. Devillers, L. Vidrascu, and L. Lamel, “Challenges in real-life emotion annotation and machine learning based detection,” *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[59] J. Diehl, D. Watson, L. Bennetto, J. McDonough, and C. Gunlogson, “An acoustic analysis of prosody in high-functioning autism,” *Applied Psycholinguistics*, vol. 30, no. 03, pp. 385–404, 2009.

[60] S. Dimidjian, C. R. Martell, and A. Christensen, *Clinical Handbook of Couple Therapy*, 4th ed. The Guilford Press, 2008, ch. Integrative behavioral couple therapy, pp. 73–106.

[61] D. Donoho, V. Stodden, and Y. Tsaig, “Sparselab,” *Retrieved Febuary*, vol. 22, p. 2009, 2009.

[62] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, “Emotional speech: Towards a new generation of databases,” *Speech Communication*, vol. 40, no. 1-2, pp. 33–60, 2003.

[63] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, “The HUMAINE Database: Addressing the collection and annotation of naturalistic and induced emotional data,” in *Affective Computing and Intelligent Interaction*, Lisbon, Portugal, 2007, pp. 488–500.

[64] J. Duchateau and L. Cleuren, “Automatic assessment of children’s reading level,” *Proc. Interspeech*, 2007.

[65] J. Duchateau, M. Wigham, K. Demuynck, and H. Hamme, “A flexible recogniser architecture in a reading tutor for children,” in *Speech Recognition and Intrinsic Variation Workshop*, 2006.

[66] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *Annals of statistics*, vol. 32, no. 2, pp. 407–451, 2004.

[67] S. Eikeseth, “Outcome of comprehensive psycho-educational interventions for young children with autism,” *Research in developmental disabilities*, vol. 30, no. 1, pp. 158–178, 2009.

[68] P. Ekman and W. Friesen, “Facial action coding system: A technique for the measurement of facial movement,” 1978.

[69] M. Eskenazi, “An overview of spoken language technology for education,” *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[70] M. Eskenazi, J. Mostow, and D. Graff, “The CMU kids speech corpus,” *Corpus of children’s read speech digitized and transcribed on two CD-ROMs, with assistance from Multicom Research and David Graff. Published by the Linguistic Data Consortium, University of Pennsylvania*, 1997.

[71] E. Ettelaie, P. G. Georgiou, and S. S. Narayanan, “Cross-lingual dialog model for speech to speech translation,” in *Proc. Interspeech*, 2006.

[72] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - The Munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.

[73] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "LIBLINEAR: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[74] W. Fay and A. Schuler, *Emerging language in autistic children.* Univ Park Pr, 1980.

[75] S. J. Fredman, D. H. Baucom, D. J. Miklowitz, and S. E. Stanton, "Observed emotional involvement and overinvolvement in families of patients with bipolar disorder," *Journal of Family Psychology*, vol. 22, no. 1, pp. 71–79, Feb. 2008.

[76] P. G. Georgiou, M. P. Black, A. C. Lammert, B. R. Baucom, and S. S. Narayanan, ""That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Affective Computing and Intelligent Interaction*, Memphis, TN, USA, 2011.

[77] P. K. Ghosh, A. Tsiartas, and S. S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio, Speech, and Language Processing*, 2010, accepted.

[78] J. Gibson, A. Katsamanis, M. P. Black, and S. S. Narayanan, "Automatic identification of salient acoustic instances in couples' behavioral interactions using Diverse Density Support Vector Machines," in *Proc. Interspeech*, Florence, Italy, 2011.

[79] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[80] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*, 2002.

[81] G. C. Gonzaga, B. Campos, and T. Bradbury, "Similarity, convergence, and relationship satisfaction in dating and married couples," *Journal of Personality and Social Psychology*, vol. 93, no. 1, pp. 34–48, 2007.

[82] K. Gotham, S. Risi, A. Pickles, and C. Lord, "The Autism Diagnostic Observation Schedule: Revised algorithms for improved diagnostic validity," *Journal of Autism and Developmental Disorders*, vol. 37, no. 4, pp. 613–627, 2007.

[83] J. Gottman, H. Markman, and C. Notarius, "The topography of marital conflict: A sequential analysis of verbal and nonverbal behavior," *Journal of Marriage and the Family*, vol. 39, no. 3, pp. 461–477, 1977.

[84] J. M. Gottman and L. J. Krokoff, "Marital interaction and satisfaction: A longitudinal view," *Journal of Consulting and Clinical Psychology*, vol. 57, no. 1, pp. 47–52, 1989.

[85] M. Greene and A. Oliva, "Recognition of natural scenes from global properties: Seeing the forest without representing the trees," *Cognitive Psychology*, vol. 58, no. 2, pp. 137–176, 2009.

[86] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[87] A. Hagen and B. Pellom, "A Multi-layered lexical-tree based token passing architecture for efficient recognition of subword speech units," in *2nd Language & Technology Conference, Poznan, Poland*, 2005.

[88] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, St. Thomas, USA*, 2003.

[89] ——, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, 2007.

[90] A. Hagen, B. Pellom, S. Van Vuuren, and R. Cole, "Advances in children's speech recognition within an interactive literacy tutor," in *Proceedings of HLT-NAACL 2004*. Association for Computational Linguistics, 2004, pp. 25–28.

[91] A. Hagen and B. Pellom, "Data driven subword unit modeling for speech recognition and its application to interactive reading tutors," in *Ninth European Conference on Speech Communication and Technology*. ISCA, 2005.

[92] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1590–1601, 2008.

[93] C. Heavey, D. Gill, and A. Christensen, *Couples interaction rating system 2 (CIRS2)*, University of California, Los Angeles, 2002.

[94] P. A. Heeman, R. Lunsford, E. Selfridge, L. Black, and J. van Santen, "Autism and interactional aspects of dialogue," in *SIGdial Meeting on Discourse and Dialogue*, Tokyo, Japan, Sept. 2010.

[95] M. Heritage, "Formative assessment: What do teachers need to know and do?" *Phi Delta Kappan*, vol. 89, no. 2, pp. 140–145, 2007.

[96] R. E. Heyman, "Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations." *Psychological Assessment*, vol. 13, no. 1, pp. 5–35, 2001.

[97] H. Hops, T. A. Wills, G. R. Patterson, and R. L. Weiss, "Marital Interaction Coding System," University of Oregon, Eugene, Oregon, USA, Tech. Rep., Dec. 1971.

[98] T. Joachims, “Text categorization with Support Vector Machines: Learning with many relevant features,” in *European Conference on Machine Learning*, Chemnitz, Germany, 1998, pp. 137–142.

[99] J. Jones and A. Christensen, *Couples interaction study: Social support interaction rating system*, University of California, Los Angeles, 1998.

[100] D. Jurafsky, R. Ranganath, and D. McFarland, “Extracting social meaning: Identifying interactional style in spoken conversation,” in *Human Language Technologies*, Boulder, CO, USA, 2009, pp. 638–646.

[101] P. Juslin and K. Scherer, “Vocal expression of affect,” *The new handbook of methods in nonverbal behavior research*, pp. 65–135, 2005.

[102] L. Kanner, “Autistic disturbances of affective contact,” *Nervous Child*, vol. 2, pp. 217–250, 1943.

[103] B. R. Karney and T. N. Bradbury, “The longitudinal course of marital quality and stability: A review of theory, methods, and research.” *Psychological Bulletin*, vol. 118, no. 1, pp. 3–34, 1995.

[104] A. Katsamanis, M. P. Black, P. G. Georgiou, L. Goldstein, and S. S. Narayanan, “SailAlign: Robust long speech-text alignment,” in *Very-Large-Scale Phonetics Workshop*, Philadelphia, PA, USA, Jan. 2011.

[105] A. Katsamanis, J. Gibson, M. P. Black, and S. S. Narayanan, “Multiple instance learning for classification of human behavior observations,” in *Affective Computing and Intelligent Interaction*, Memphis, TN, USA, 2011.

[106] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, “TBALL data collection: The making of a young children’s speech corpus,” in *Proc. Interspeech*, 2005.

[107] D. Keen, “The use of non-verbal repair strategies by children with autism,” *Research in Developmental Disabilities*, vol. 26, no. 3, pp. 243–254, 2005.

[108] P. K. Kerig and D. H. Baucom, Eds., *Couple Observational Coding Systems.* Mahwah, NJ, USA: Lawrence Erlbaum, 2004.

[109] C.-C. Lee, M. P. Black, A. Katsamanis, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, “Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples,” in *Proc. Interspeech*, 2010.

[110] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, P. G. Georgiou, and S. S. Narayanan, “An analysis of PCA-based vocal entrainment measures in married couples’ affective spoken interactions,” in *Proc. Interspeech*, Florence, Italy, 2011.

[111] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” in *Proc. Interspeech*, 2009.

[112] C.-C. Lee and S. Narayanan, "Predicting interruptions in dyadic spoken interactions," in *IEEE ICASSP*, 2010.

[113] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[114] K. Lee, A. Hagen, N. Romanyshyn, S. Martin, and B. Pellom, "Analysis and detection of reading miscues for interactive literacy tutors," in *Proceedings of the 20th international conference on Computational Linguistics.* Association for Computational Linguistics, 2004, p. 1254.

[115] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, pp. 1455–1468, 1999.

[116] H. Levene, "Robust tests for equality of variances," in *Contributions to Probability and Statistics*, I. Olkin, Ed. Palo Alto, CA: Stanford University Press, 1960, pp. 278–292.

[117] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman *et al.*, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, vol. 4, 2007.

[118] C. Lord, S. Risi, L. Lambrecht, E. Cook, B. Leventhal, P. DiLavore, A. Pickles, and M. Rutter, "The Autism Diagnostic Observation ScheduleGeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of Autism and Developmental Disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[119] G. Margolin, E. Gordis, and P. Oliver, "Links between marital and parent–child interactions: Moderating role of husband-to-wife aggression," *Development and psychopathology*, vol. 16, no. 03, pp. 753–771, 2004.

[120] G. Margolin, P. Oliver, E. Gordis, H. O'Hearn, A. Medina, C. Ghosh, and L. Morland, "The nuts and bolts of behavioral observation of marital and family interaction," *Clinical Child and Family Psychology Review*, vol. 1, no. 4, pp. 195–213, 1998.

[121] J. McCann and S. Peppe, "Prosody in autism spectrum disorders: a critical review," *International Journal of Language & Communication Disorders*, vol. 38, no. 4, pp. 325–350, 2003.

[122] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, June 1947.

[123] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences.* Pearson Prentice Hall, 2007, ch. 7.8: Estimation of the Difference Between Two Population Proportions, pp. 302–303.

[124] *"behavior"*, Merriam-Webster Online Dictionary Std., Merriam-Webster Online Dictionary, July 2010. [Online]. Available: http://www.merriam-webster.com

[125] *"subjective"*, Merriam-Webster Online Dictionary Std., Merriam-Webster Online Dictionary, July 2010. [Online]. Available: http://www.merriam-webster.com

[126] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[127] J. Miller and A. Iglesias, "Systematic Analysis of Language Transcripts (SALT)," Language Analysis Lab, University of Wisconsin-Madison, 2006, (Version 9) [Computer Software].

[128] L. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010.

[129] L. Morency, C. Sidner, C. Lee, and T. Darrell, "Head gestures for perceptual interfaces: The role of context in improving recognition," *Artificial Intelligence*, vol. 171, no. 8-9, pp. 568–585, 2007.

[130] P. Moreno, C. Joerg, J.-M. van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Proc. ICSLP*, 1998.

[131] J. Mostow and J. Beck, "When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor That Listens1," *Scale-up in Education: Issues in practice*, p. 183, 2006.

[132] J. Mostow, A. G. Hauptmann, L. L. Chase, and S. Roth, "Towards a reading coach that listens: Automated detection of oral reading errors," in *Proc. Nat. Conf. on Artifical Intelligence.* John Wiley & Sons, 1993, pp. 392–392.

[133] J. Mostow, S. F. Roth, E. G. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proc. of Nat. Conf. on Aritifical Intelligence*, 1994.

[134] E. Mower, M. P. Black, E. Flores, M. E. Williams, and S. S. Narayanan, "Rachel: Design of an emotionally targeted interactive agent for children with autism," in *ICME*, 2011.

[135] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotional profiles," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[136] K. Murray, "A study of automatic pitch tracker doubling/halving "errors"," in *SIGdial Workshop on Discourse and Dialogue*, vol. 16, Aalborg, Denmark, 2001.

[137] H. Nait-Charif and S. McKenna, "Activity summarisation and fall detection in a supportive home environment," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*, vol. 4, 2004.

[138] National Center on Birth Defects and Developmental Disabilities. (2010, March) Autism Spectrum Disorders (ASDs). Centers for Disease Control and Prevention. [Online]. Available: http://www.cdc.gov/ncbddd/autism/

[139] National Reading Panel, "Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction," National Institute for Child Health and Human Development, National Institute of Health, Washington, D.C., Tech. Rep. 00-4769, 2000.

[140] M. O'Brien, R. S. John, G. Margolin, and O. Erel, "Reliability and diagnostic efficacy of parents' reports regarding children's exposure to marital aggression," *Violence and Victims*, vol. 9, no. 1, pp. 45–62, 1994.

[141] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[142] D. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, p. 13354, 2010.

[143] S. Otaiba and J. Torgesen, "Effects from intensive standardized kindergarten and first-grade interventions for the prevention of reading difficulties," *Handbook of response to intervention*, pp. 212–222, 2007.

[144] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human computing and machine understanding of human behavior: A survey," *Artifical Intelligence for Human Computing*, pp. 47–71, 2007.

[145] J. Paratore and R. McCormack, *Classroom literacy assessment: Making sense of what students know and do.* New York, NY: The Guilford Press, 2007.

[146] A. Pentland, "Socially aware computation and communication," in *Proceedings of the 7th international conference on Multimodal interfaces.* ACM, 2005, p. 199.

[147] P. Price, J. Tepperman, M. Iseli, T. Duong, M. Black, S. Wang, C. K. Boscardin, M. Heritage, P. David Pearson, S. Narayanan, and A. Alwan, "Assessment of emerging reading skills in young native speakers and language learners," *Speech Communication*, vol. 51, no. 10, pp. 968–984, 2009.

[148] R. Ranganath, D. Jurafsky, and D. McFarland, "It's not you, it's me: Detecting flirting and its misperception in speed-dates," in *EMNLP*, 2009.

[149] P. Rao, D. Beidel, and M. Murray, "Social skills interventions for children with Asperger's syndrome or high-functioning autism: A review and recommendations," *Journal of autism and developmental disorders*, vol. 38, no. 2, pp. 353–361, 2008.

[150] V. Rozgić, B. Xiao, A. Katsamanis, B. Baucom, P. G. Georgiou, and S. S. Narayanan, "Estimation of ordinal approach-avoidance labels in dyadic interactions: Ordinal logistic regression approach," in *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, Prague, Czech Republic, 2011, pp. 2368–2371.

[151] ——, "A new multichannel multimodal dyadic interaction database," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 1982–1985.

[152] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and L. Kessous, "The relevance of feature type for automatic classification of emotional user states: Low level descriptors and functionals," in *Proc. Interspeech*, 2007.

[153] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Proc. Interspeech*, 2009.

[154] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The Interspeech 2010 paralinguistic challenge," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 2794–2797.

[155] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?" in *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, Las Vegas, NV, USA, 2008, pp. 4501–4504.

[156] B. Schuller, M. Wöllmer, F. Eyben, and G. Rigoll, "Prosodic, spectral or voice quality? Feature type relevance for the discrimination of emotion pairs," in *The Role of Prosody in Affective Speech*, S. Hancil, Ed. Berlin, Germany: Peter Lang Publishing Group, 2009, pp. 285–307.

[157] M. Sevier, K. Eldridge, J. Jones, B. D. Doss, and A. Christensen, "Observed communication and associations with satisfaction during traditional and integrative behavioral couple therapy," *Behavior Therapy*, vol. 39, no. 2, pp. 137–150, 2008.

[158] M. Sevier, L. E. Simpson, and A. Christensen, *Demand/withdraw interaction coding*, ser. Couple observational coding systems. Mahwah, NJ, USA: Lawrence Erlbaum, 2004, pp. 159–172.

[159] K. Shobaki, J. P. Hosom, and R. A. Cole, "The OGI kids' speech corpus and recognizers," in *Sixth International Conference on Spoken Language Processing.* Citeseer, 2000.

[160] V. Shoham, M. J. Rohrbaugh, T. R. Stickle, and T. Jacob, "Demand-withdraw couple interaction moderates retention in cognitive-behavioral versus family-systems treatments for alcoholism." *Journal of Family Psychology*, vol. 12, no. 4, pp. 557–577, 1998.

[161] E. E. Shriberg, “Preliminaries to a theory of speech disfluencies,” Ph.D. dissertation, Citeseer, 1994.

[162] J. Tepperman, M. P. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan, and S. Narayanan, “A Bayesian network classifier for word-level reading assessment,” *Proceedings of ICSLP, Antwerp, Belgium*, 2007.

[163] J. Tepperman, M. Gerosa, and S. Narayanan, “A generative model for scoring children’s reading comprehension,” *Proc. of the Workshop on Child, Computer, and Interaction*, 2008.

[164] J. Tepperman, S. Lee, A. Alwan, and S. Narayanan, “A generative student model for scoring word reading skills,” *IEEE Transactions on Audio, Speech, and Language Processing*, no. Accepted, 2010.

[165] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, “Pronunciation verification of children’s speech for automatic literacy assessment,” in *Ninth International Conference on Spoken Language Processing*, 2006.

[166] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[167] S. E. Tranter and D. A. Reynolds, “An overview of automatic speaker diarization systems,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[168] H. Traunmüller and A. Eriksson, “The frequency range of the voice fundamental in the speech of male and female adults,” Linguistics Department, University of Stockholm, Tech. Rep., 1994.

[169] S. Tuchschmid, M. Bajka, and M. Harders, “Comparing Automatic Simulator Assessment with Expert Assessment of Virtual Surgical Procedures,” *Biomedical Simulation*, pp. 181–191, 2010.

[170] M. Valstar and M. Pantic, “Fully automatic facial action unit detection and temporal analysis,” 2006.

[171] J. P. H. van Santen, E. T. Prud’hommeaux, and L. M. Black, “Automated assessment of prosody production,” *Speech communication*, vol. 51, no. 11, pp. 1082–1097, 2009.

[172] J. P. H. van Santen, E. T. Prud’hommeaux, L. M. Black, and M. Mitchell, “Computational prosodic markers for autism,” *Autism*, p. 1362361310363281v1, 2010.

[173] M. Venezia, D. S. Messinger, D. Thorp, and P. Mundy, “The development of anticipatory smiling,” *Infancy*, vol. 6, no. 3, pp. 397–406, 2004.

[174] A. Vinciarelli, M. Pantic, and H. Bourlard, “Social signal processing: Survey of an emerging domain,” *Image and Vision Computing*, vol. 27, pp. 1743–1759, 2009.

[175] S. Wang, P. Price, M. Heritage, and A. Alwan, “Automatic evaluation of children’s performance on an English syllable blending task,” in *Proc. SLaTE*, 2007.

[176] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, “Toward Practical Smile Detection,” *IEEE transactions on pattern analysis and machine intelligence*, pp. 2106–2111, 2009.

[177] S. M. Williams, D. Nix, and P. Fairweather, “Using speech recognition technology to enhance literacy instruction for emerging readers,” *Proceedings of ICLS 2000*, p. 115, 2000.

[178] K. J. Williams-Baucom, D. C. Atkins, M. I. A. Sevier, K. A. Eldridge, and A. Christensen, ““You” and “I” need to talk about “us”: Linguistic patterns in marital interactions,” *Personal Relationships*, vol. 17, no. 1, pp. 41–56, 2010.

[179] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[180] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan, “Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling,” in *Interspeech*, 2010.

[181] S. Yildirim, S. Narayanan, and A. Potamianos, “Detecting emotional state of a child in a conversational computer game,” *Computer Speech & Language*, 2010.

[182] H. You, A. Alwan, A. Kazemzadeh, and S. Narayanan, “Pronunciation variations of Spanish-accented English spoken by young children,” in *Proc. Interspeech*, 2005.

[183] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, “The HTK book (for HTK version 3.4),” *Cambridge University Engineering Department*, 2006.