Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 1–21

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features ☆

Matthew P. Black [a,*], Athanasios Katsamanis [a], Brian R. Baucom [b], Chi-Chun Lee [a], Adam C. Lammert [a], Andrew Christensen [c], Panayiotis G. Georgiou [a]
Shrikanth S. Narayanan [a,b]

[a] *Signal Analysis & Interpretation Laboratory (SAIL), University of Southern California, 3710 McClintock Ave., Los Angeles, CA 90089, USA*[1]
[b] *Department of Psychology, University of Southern California (USC), 3620 McClintock Ave., Los Angeles, CA 90089, USA*
[c] *Department of Psychology, University of California, Los Angeles (UCLA), 1285 Franz Hall, Los Angeles, CA 90095, USA*

Available online 16 December 2011

Abstract

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 this work, 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, an improvement of 3.95% absolute (5.63% relative) over our previously reported best results. We compare performance for the various factors: across codes, gender, classifier type, and feature type.

Keywords: Behavioral signal processing (BSP); Couple therapy; Dyadic interaction; Human behavior analysis; Prosody; Emotion recognition

1. Introduction

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 (O'Brien et al., 1994).

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 (Baucom et al., 2007), bi-polar disorder (Fredman et al., 2008), anxiety (Beck et al., 2006), schizophrenia (Brüne et al., 2008), autism (Keen, 2005), alcoholism (Shoham et al., 1998), domestic aggression (Margolin et al., 2004), and marital distress (Heyman, 2001). In each of these areas, observational research has identified behaviors

☆ "The original version of this paper was selected as one of the best papers from Interspeech 2010. It is presented here in revised form following additional peer review".

* Corresponding author. Tel.: +1 213 740 6432; fax: +1 213 740 4651. *E-mail address:* matthepb@usc.edu (M.P. Black).

[1] The SAIL homepage is http://sail.usc.edu.

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.12.003

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 (Karney and Bradbury, 1995; Gonzaga et al., 2007), 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 (Margolin et al., 1998). This manual coding is a costly and challenging process. First, a detailed coding manual must be designed, which can be a complex iterative task (Kerig and Baucom, 2004).

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 (Kerig and Baucom, 2004). 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., Hops et al., 1971). 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 of 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 (Schuller et al., 2007), 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 work 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 marital therapy).

Significant work related to BSP has concentrated on extracting human-centered information from audio-video signals, including social cues (Vinciarelli et al., 2009), affect and emotions (Lee and Narayanan, 2005; Grimm et al., 2007; Schuller et al., 2009a; Yildirim et al., 2010), and intent (Jurafsky et al., 2009). 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 (Campbell, 2000; Douglas-Cowie et al., 2003, 2007; Devillers et al., 2005; Devillers and Campbell, 2011; Burkhardt et al., 2009).

This paper builds upon some of our recent work in applying the basic ideas of BSP using the Couple Therapy corpus (Christensen et al., 2004), discussed in detail in Section 2. 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 (Black et al., 2010), 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 (Lee et al., 2010), we developed quantitative methods to model prosodic *entrainment* behavior between the spouses; couples rated as behaving more positive were found to have statistically 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 paper represents an extension of Black et al. (2010), in which we analyze the same corpus. In this work, 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 (Black et al., 2010).

Section 2 describes the Couple Therapy corpus, and Section 3 provides a methodological overview. We explain how we pre-processed the data in Section 4. Section 5 discusses the acoustic features we extracted to model the spouses' behavior, while Section 6 describes the learning methods and algorithms used to predict the spouses' behavioral codes. The results are presented and discussed in Section 7, and the conclusions and intended future work are provided in Section 8.

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 (Christensen et al., 2004). The main purpose was to test the efficacy of integrative behavioral couple therapy (IBCT, Christensen et al., 1995) versus traditional behavioral couple therapy (e.g., Baucom et al., 1998) 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 (Christensen et al., 2004, 2006, 2010; Baucom et al., 2009). 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 (704 × 480 pixels, 29.97 fps) and a single channel of far-field audio recorded from the videocamera microphone (16 kHz, 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 (Sevier et al., 2008). The Social Support Interaction Rating System (SSIRS) measured both the emotional content of the interaction as well as the topic of conversation (Jones and Christensen, 1998). It consisted of 19 questions ("codes") across four categories: affectivity, dominance/

Table 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).

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 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.

Code	Code correlation					Spouse	Agreement
	acc	bla	pos	neg	sad	Correlation	
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

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 (Heavey et al., 2002). All 32 codes had written guidelines and were on an integer scale from 1 ("none/not at all") to 9 ("a lot"). Table 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 2 and explained in further detail in (Sevier et al., 2004). 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 (Sevier et al., 2008). 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 (Atkins et al., 2005; Baucom et al., 2009; Williams-Baucom et al., 2010). 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. Methodology overview

The Couple Therapy corpus provides a unique opportunity to test BSP methods/algorithms on data collected in an

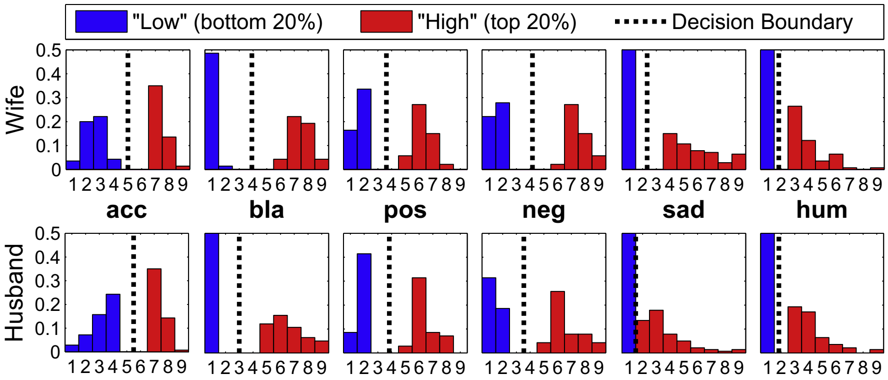

Fig. 1. Normalized histograms of the extreme code scores for the wife (top) and husband (bottom). The "low" scores are in the bottom 20%, and the "high" scores are in the top 20%. The decision boundary was used to compute an upper-bound for automatic performance.

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 may be present even in corpora collected with high-quality recording equipment, consistent sensor locations, and controlled acoustic/visual environmental conditions (e.g., Rozgić et al., 2010). We believe that analyzing this existing large corpus offers a veritable testbed for this domain of BSP research.

In this paper, 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 3.1 describes the subset of codes we analyzed, and Section 3.2 explains the classification set-up for all experiments. Section 3.3 provides an overview of our methodology: data pre-processing, acoustic feature extraction, and supervised learning of the behavioral codes. Sections 4–6 provide more detailed descriptions of these three components, respectively.

3.1. Codes of interest

For clarity and to make the results comparable to our previous work (Black et al., 2010), 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"). Appendix A provides the written guidelines for the six codes. It should be noted that each code measures how much that particular code occurred, *not* 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 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 paper had inter-evaluator agreement that ranged from 0.4 to 0.7.

We see in Table 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 2 that the two spouses' behaviors were positively correlated for all six codes; this suggests that, on average, the interacting spouses displayed similar behaviors.

3.2. Classification task formulation

As described in Section 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 Fig. 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 (Black et al., 2010) and in related work (Jurafsky et al., 2009; Ranganath et al., 2009). This is a good starting

Table 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.

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

point in trying to learn these subtle high-level behavioral codes.

As shown in Fig. 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 Fig. 1, and the upper-bounds in code performance for the wife and husband are listed in Table 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 Fig. 1).

3.3. *Classification system overview*

See Fig. 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 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* (Katsamanis et al., 2011a), 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.

For this paper, 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

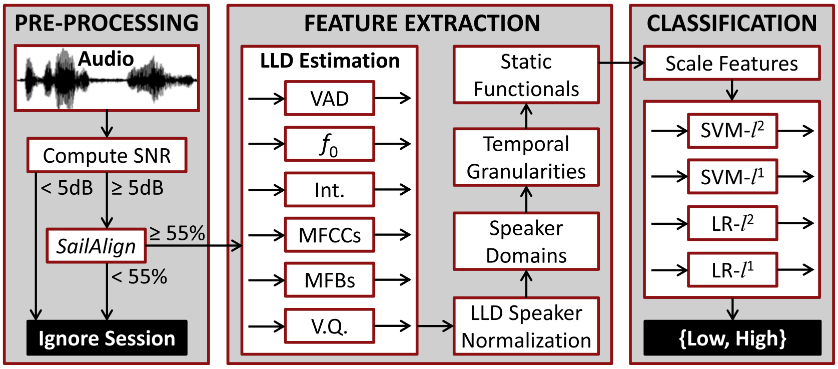

Fig. 2. A system block diagram, illustrating the methodology taken in this paper, from pre-processing the data and extracting acoustic features to classifying extreme instances of a particular code as low/high.

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 (Christensen and Heavey, 1990). 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.

4. Data pre-processing

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 (Ghosh et al., 2010). 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 2).

We used the VAD output to estimate the average SNR of each session's audio file using Eq. (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):

$$\mathrm{SNR(dB)} = 10\log_{10}\frac{\frac{1}{|i\in S|}\sum_{i\in S}A_i^2}{\frac{1}{|i\notin S|}\sum_{i\notin S}A_i^2}, \tag{1}$$

Fig. 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.

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

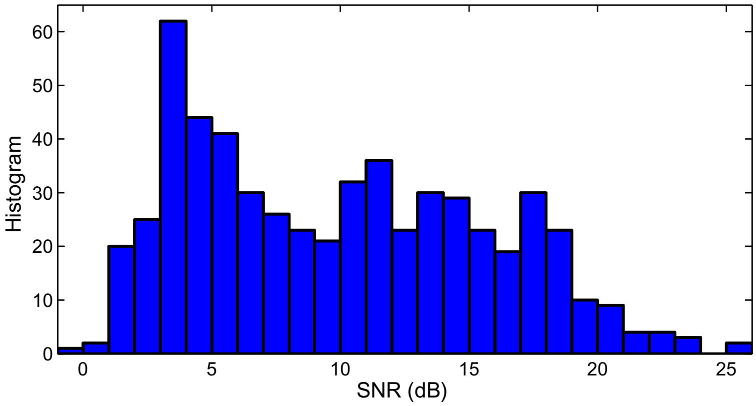

Fig. 3. A histogram of the estimated average signal-to-noise ratio (SNR) for each of the 569 coded sessions, computed using Eq. (1).

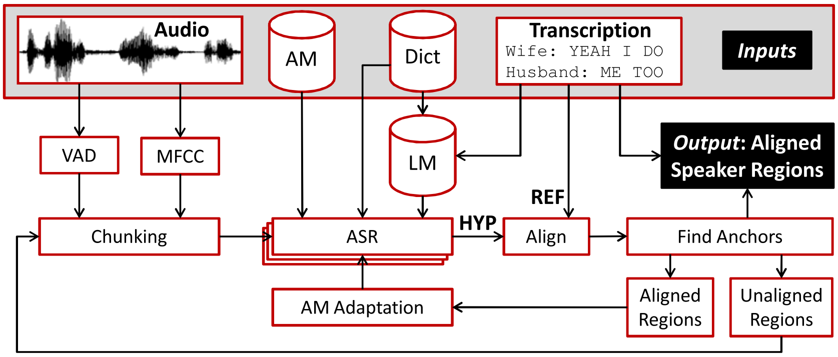

Fig. 4. Block diagram of the "hybrid" manual/automatic speaker segmentation procedure, implemented using *SailAlign*. See Section 4.2 and Katsamanis et al. (2011a) for details.

conditions). However, manually segmenting a corpus of this size was not practical and is not scalable.

In this paper and in our previous work (Black et al., 2010), we took a unique "hybrid" manual/automatic speaker segmentation approach that exploited the available transcriptions with speaker labels. We implemented a recursive ASR-based procedure to align the transcription with the corresponding audio using *SailAlign* (Katsamanis et al., 2011a), open-source software we developed as part of this work. The iterative algorithm was based on the work by Moreno et al. (1998), with the extension that aligned portions of the audio were used to adapt the acoustic models at each iteration.

Fig. 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 by the VAD that split the MFCC feature vector sequence 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. See Katsamanis et al. (2011a) 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. Fig. 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 paper, we ignored 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.

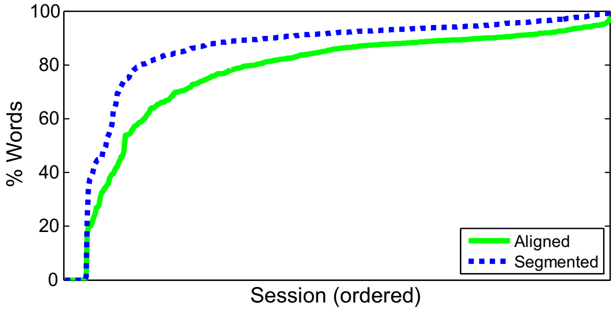

Fig. 5. The percentage of words *aligned* using *SailAlign* and the percentage of words that were subsequently *segmented* into single speaker regions for the 415 sessions with SNR greater than 5 dB.

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.

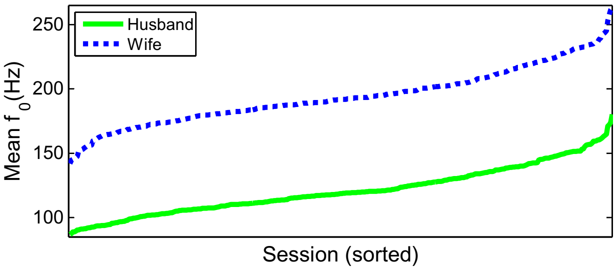

Fig. 6. The ordered mean fundamental frequency (f_0) estimates for the wife and husband in each of the 372 coded sessions that met the SNR and speaker segmentation criteria.

Table 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 (Traunmüller and Eriksson, 1994, p.3).

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	–

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 (Traunmüller and Eriksson, 1994). We estimated f_0 for each session (see Section 5) and computed f_0 statistics for the husband and wife across the speaker-homogeneous regions.

Fig. 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 4 shows that the average f_0 statistics are similar to the ones reported in (Traunmüller and Eriksson, 1994, p. 3), 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 (Black et al., 2010), 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.

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., Juslin and Scherer, 2005; Cowie, 2009), including in those related to marital interactions (Gottman et al., 1977; Gottman and Krokoff, 1989; Baucom et al., 2009). Affect/emotion are discussed as critical components to communication and are oftentimes conveyed vocally.

In our previous paper (Black et al., 2010), 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 (Lee and Narayanan, 2005; Grimm et al., 2007; Schuller et al., 2007, 2009a; Lee et al., 2009; Ranganath et al., 2009; Yildirim et al., 2010).

We examined an expanded set of features in this paper 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 7, Fig. 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, the session-level features were computed 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 5 lists each of these components and Sections 5.1–5.4 provide further details.

5.1. Low-level descriptors

In this work, we refer to low-level descriptors (LLDs) as feature streams that are estimated/extracted at fine temporal resolutions (e.g., every 10 ms). Table 5 lists each of the LLDs we selected for this paper, based on our previous work (Black et al., 2010) and on the 2009 Interspeech Emotion Challenge (Schuller et al., 2009a) and 2010 Interspeech Paralinguistic Challenge (Schuller et al., 2010).

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, developed for a speech production modeling toolkit.[2] Therefore, this speaking rate LLD was at the word-level and only applicable to words that were aligned with *SailAlign* (see Section 4.2). Another LLD we extracted directly from the alignment results (when available) were the interturn 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

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

Table 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.

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*, 1st percentile*, 99th percentile*, 99th – 1st percentile*, skewness, kurtosis, minimum position, maximum position, lower quartile, upper quartile, interquartile range, linear approximation slope

shimmer. f_0 and intensity were extracted with Praat (Boersma, 2001), and the other LLDs were extracted with openSMILE (Eyben et al., 2010). 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 (Juslin and Scherer, 2005) and emotion recognition research (Grimm et al., 2007; Bulut and Narayanan, 2008; Busso et al.,2009a,b; Lee et al., 2009; Yildirim et al., 2010). 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 (Boersma, 2001). 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 (Murray, 2001; Coy and Barker, 2007; Chen et al., 2004). 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 and halving/doubling the f_0 signal toward the mean f_0 value of the speaker.

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 Fig. 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, Eq. (2), subtracts the mean $f_0(\mu_{f_0})$ of the speaker (wife, husband, or unknown) for each frame. The second method, Eq. (3), performs a similar transformation on a logarithmic scale, since this may be more perceptually motivated (de Cheveigné and Kawahara, 2002). 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."

$$\bar{f}_{0_{\text{lin}}} = f_0 - \mu_{f_0}, \tag{2}$$

$$\bar{f}_{0_{\log}} = \log_2\left(\frac{f_0}{\mu_{f_0}}\right). \tag{3}$$

The computation of intensity for an audio signal is more straightforward 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). Eq. (4) shows how we normalized each

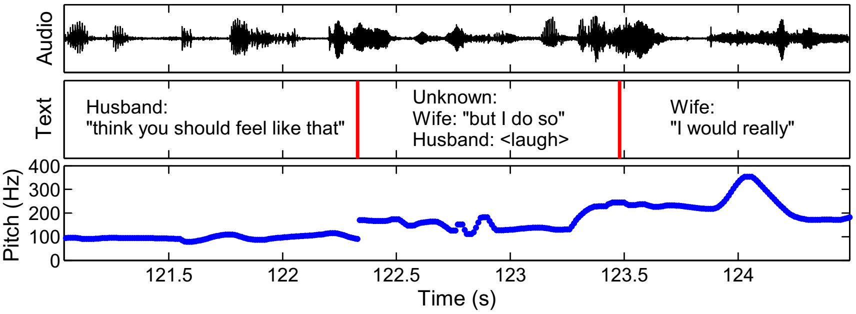

Fig. 7. Example of the speaker segmentation and processed f_0 signal. In this particular example, the middle portion (labeled "Unknown") was unable to be automatically segmented due to overlapped speech (the husband was laughing while the wife was speaking).

frame-level intensity value, where the μ_{int} values were the mean intensity of the speaker during speech regions, computed across the whole session:

$$\text{int}_n = \frac{\text{int}}{\mu_{\text{int}}}. \tag{4}$$

We used openSMILE to extract spectral and voice quality features using the same parameter settings as the 2010 Interspeech Paralinguistic Challenge (Schuller et al., 2010). 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 Eq. (5), where the $\mu_{\text{MFCC}[i]}$ values were the mean MFCC of the ith coefficient of the speaker, computed across the whole session:

$$\text{MFCC}_n[i] = \text{MFCC}[i] - \mu_{\text{MFCC}[i]}, \quad i = 0, \ldots, 14. \tag{5}$$

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 (Gobl and Chasaide, 2003), although most engineering studies have found they are often less discriminative than the more traditional prosodic and spectral features (e.g., Schuller et al., 2009b), 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 (Eyben et al., 2010).

5.2. Speaker domains

For all the LLDs described in Section 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 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. 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 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 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.

Table 6
A depiction of which speech regions were included in the five speaker domains, depending on which spouse was being rated.

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		✓

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 (Black et al., 2010).

For the *halves* granularity, we split each LLD stream into two halves and computed the difference in functionals (see Section 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 5.4. The hierarchical temporal granularity was based on the work by Schuller et al. (2008) 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.

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 5 for the full list of 14 functionals that we selected. Note that the 1st percentile, 99th percentile, and 99th − 1st 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 (Schuller et al., 2008). These six basic functionals are starred (*) in Table 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.

6. Prediction of behavioral codes

Given that there were 372 sessions that were deemed acceptable after pre-processing the corpus (see Section 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 (Joachims, 1998) and emotion recognition (Batliner et al., 2011).

In our previous paper (Black et al., 2010), 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 work, 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 paper, 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 Eqs. (6)–(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 (Eqs. 6 and 8) are more commonly used, the l^1-regularized classifiers (Eqs. 7 and 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 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.

$$\hat{\mathbf{w}}_{\mathrm{SVM}-l^2} = \min_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m} \max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)^2 \right), \quad (6)$$

$$\hat{\mathbf{w}}_{\mathrm{SVM}-l^1} = \min_{\mathbf{w}} \left(\|\mathbf{w}\|_1 + C\sum_{i=1}^{m} \max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)^2 \right), \quad (7)$$

$$\hat{\mathbf{w}}_{\mathrm{LR}-l^2} = \min_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m} \log(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}) \right), \quad (8)$$

$$\hat{\mathbf{w}}_{\mathrm{LR}-l^1} = \min_{\mathbf{w}} \left(\|\mathbf{w}\|_1 + C\sum_{i=1}^{m} \log(1 + e^{-y_i\mathbf{w}^T\mathbf{x}_i}) \right). \quad (9)$$

We used the implementations in LIBLINEAR for all four classifiers (Fan et al., 2008). Note that the primal forms of the loss functions are written in Eqs. (6)–(8) for clarity. In practice, the dual forms were faster to train; see Fan et al. (2008) 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, 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:

$$\hat{y} = \mathrm{sgn}(\hat{\mathbf{w}}^T\mathbf{x}). \quad (10)$$

7. Results & discussion

Table 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). We see from Table 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 8 discusses ongoing and future work to improve upon the results achieved in this paper.

Table 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 (Eqs. (6)–(9)). Baseline chance performance was 50%, and an upper-bound was estimated using individual human evaluator scores (Section 3.2). The bold values signify the highest accuracy for a particular code and gender.

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

We used two statistical tests to determine if the differences in performance were statistically significant: a one-sided McNemar's test for "paired" instances (McNemar, 1947), 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 (Mendenhall and Sincich, 2007), which occurred when comparing different codes or different rated spouse genders.

All results shown in Table 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 (Black et al., 2010), 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 work, 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 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 ($p < 0.01$).

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 (Eqs. 7 and 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 (Christensen and Heavey, 1990); 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 Fig. 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 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 8 provides details on the number and fraction of selected features for the various feature

Table 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 5.2).

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.–0.1 s	7560	12600	132.4	164.5	206.7	0.154	0.169	0.162	0.018	0.022	0.016
	hier.–0.5 s	7560	12600	147.7	170.2	220.8	0.154	0.152	0.156	0.020	0.023	0.018
	hier.–1 s	7560	12600	165.0	197.6	251.1	0.179	0.189	0.188	0.022	0.026	0.020
	hier.–5 s	7560	12600	188.4	218.5	271.5	0.215	0.198	0.202	0.025	0.029	0.022
	hier.–10 s	7560	12600	187.8	215.2	266.1	0.208	0.201	0.189	0.025	0.029	0.021

components and sub-components ("subsets") proposed in this paper (see Section 5).

In the first row of Table 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 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 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. Fig. 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 Fig. 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.

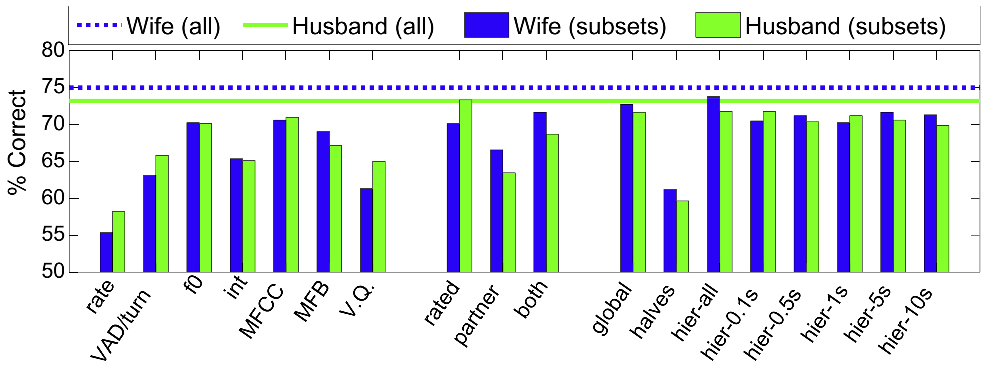

Fig. 8. Average percentage of correctly classified instances across the six codes for the wife and husband (using gender-specific models and l^2-regularized logistic regression) for single feature *subsets* and compared to the case when we used *all* the features.

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

Table 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.

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**

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 9 and Fig. 9 show the results from this final experiment for all six codes and both spouses. We see from Table 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%

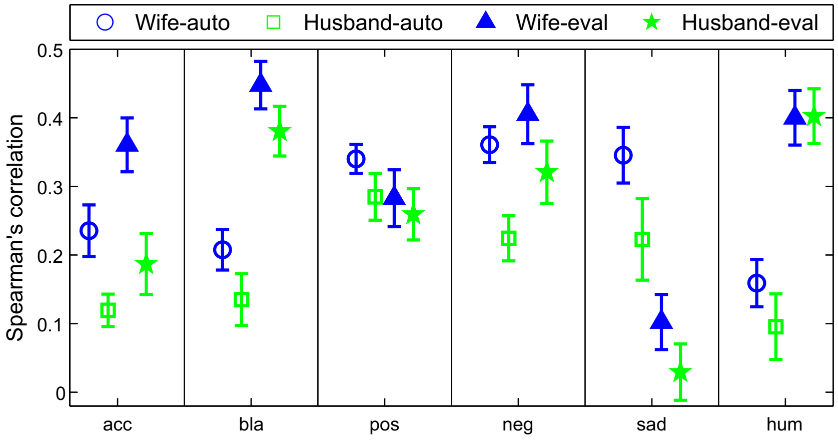

Fig. 9. Performance in ranking the instances of the middle 60% of the six behavioral codes for the automatic (auto) system, as compared to inter-evaluator agreement (eval). Each plot shows the mean and standard deviation in the Spearman's correlation.

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 9 and Fig. 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 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.

8. Conclusions & future work

In this work, 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 is 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., Batliner et al., 2011) 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 (Lee et al., 2010, 2011).

While the methods proposed in this paper focused on extracting acoustic speech cues, we are also studying the predictive power of lexical cues. Incorporating lexical features may be especially important for codes like "level of blame," where the coding manual description in Appendix A explicitly instructs evaluators to pay attention to specific language use. We are using both the manual transcriptions and automatically-generated transcriptions (through automatic speech recognition) to predict the session-level behavioral codes (Georgiou et al., 2011). Since lexical features are potentially complementary to the acoustic features used in this paper, we are also experimenting with methods to fuse these two information sources (Black et al., 2011; Katsamanis et al., 2011b).

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 (Katsamanis et al., 2011b; Gibson et al., 2011).

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 A). 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 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 paper 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., Rozgić et al., 2011). 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., Tranter and Reynolds, 2006; Han et al., 2008). 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 Appendix A, 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 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 (Rozgić et al., 2010).

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.

Acknowledgments

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, and coding the data.

Appendix A. Coding manual written guidelines

Below we provide the written guidelines of the six codes analyzed in this paper, copied from the two coding manuals. "Acceptance of other" and "blame" were from the Couples Interaction Rating System (CIRS, Heavey et al., 2002), and "global positive," "global negative," "sadness," and "use of humor" were from the Social Support Interaction Rating System (SSIRS, Jones and Christensen, 1998).

> **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 do not 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.

References

Atkins, D., Milbright, S.A., Dueck, A., Reimer, K., Christensen, A., 2005. 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.

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

Baucom, D.H., Shoham, V., Mueser, K.T., Daiuto, A.D., Stickle, T.R., 1998. Empirically supported couple and family interventions for marital distress and adult mental health problems. J. Consult. Clin. Psychol. 66, 53–88.

Baucom, B.R., Eldridge, K., Jones, J., Sevier, M., Clements, M., Markman, H., Stanley, S., Sayers, S.L., Sher, T., Christensen, A., 2007. Relative contributions of relationship distress and depression to communication patterns in couples. J. Soc. Clin. Psychol. 26, 689–707.

Baucom, B.R., Atkins, D.C., Simpson, L.E., Christensen, A., 2009. Prediction of response to treatment in a randomized clinical trial of couple therapy: A 2-year follow-up. J. Consult. Clin. Psychol. 77, 160–173.

Beck, J.G., Davila, J., Farrow, S., Grant, D., 2006. When the heat is on: Romantic partner responses influence distress in socially anxious women. Behav. Res. Ther. 44, 737–748.

Black, M.P., Katsamanis, A., Lee, C.C., Lammert, A.C., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.S. 2010. Automatic classification of married couples' behavior using audio features. In: Proc. Interspeech, Makuhari, Japan. pp. 2030–2033.

Black, M.P., Georgiou, P.G., Katsamanis, A., Baucom, B.R., Narayanan, S.S. 2011. "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.

Boersma, P.P.G., 2001. Praat, a system for doing phonetics by computer. Glot Internat. 5, 341–345.

Brüne, M., Sonntag, C., Abdel-Hamid, M., Lehmkämper, C., Juckel, G., Troisi, A., 2008. Nonverbal behavior during standardized interviews in patients with schizophrenia spectrum disorders. J. Nerv. Ment. Dis. 196, 282–288.

Bulut, M., Narayanan, S.S., 2008. On the robustness of overall F0-only modifications to the perception of emotions in speech. J. Acoust. Soc. Amer 123, 4547–4558.

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

Busso, C., Bulut, M., Lee, S., Narayanan, S.S. 2009a. Fundamental frequency analysis for speech emotion processing. In: Hancil, S. (Ed.), The Role of Prosody in Affective Speech. Peter Lang Publishing Group, Berlin, Germany, pp. 309–337.

Busso, C., Lee, S., Narayanan, S.S., 2009b. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Audio, Speech, Lang. Process. 17, 582–596.

Campbell, N. 2000. Databases of emotional speech, in: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

Chen, K., Hasegawa-Johnson, M., Cohen, A. 2004. 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. pp. 509–512.

Christensen, A., Heavey, C.L., 1990. Gender differences in marital conflict: The demand/withdraw interaction pattern. Gender Issues Contemp. Soc. 6, 113–141.

Christensen, A., Jacobson, N.S., Babcock, J.C., 1995. Integrative behavioral couple therapy. In: Jacobsen, N.S., Gurman, A.S. (Eds.), Clinical Handbook of Marital Therapy, second ed. Guilford Press, New York, pp. 31–64.

Christensen, A., Atkins, D.C., Berns, S., Wheeler, J., Baucom, D.H., Simpson, L.E., 2004. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. J. Consult. Clin. Psychol. 72, 176–191.

Christensen, A., Atkins, D.C., Yi, J., Baucom, D.H., George, W.H., 2006. Couple and individual adjustment for 2 years following a randomized clinical trial comparing traditional versus integrative behavioral couple therapy. J. Consult. Clin. Psychol. 74, 1180–1191.

Christensen, A., Atkins, D.C., Baucom, B.R., Yi, J., 2010. Marital status and satisfaction five years following a randomized clinical trial comparing traditional versus integrative behavioral couple therapy. J. Consult. Clin. Psychol. 78, 225–235.

Cowie, R., 2009. Perceiving emotion: Towards a realistic understanding of the task. Philos. Trans. Roy. Soc. B: Biological Sci. 364, 3515–3525.

Coy, A., Barker, J., 2007. An automatic speech recognition system based on the scene analysis account of auditory perception. Speech Comm. 49, 384–401.

de Cheveigné, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Amer 111, 1917–1930.

Devillers, L., Campbell, N., 2011. Special issue of computer speech and language on "affective speech in real-life interactions". Comput. Speech Lang. 25, 1–3.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: Towards a new generation of databases. Speech Comm. 40, 33–60.

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

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

Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C., 2008. LIBLINEAR: A library for large linear classification. J. Machine Learn. Res. 9, 1871–1874.

Fredman, S., Baucom, D., Miklowitz, D., Stanton, S., 2008. Observed emotional involvement and overinvolvement in families of patients with bipolar disorder. J. Fam. Psychol. 22, 71–79.

Georgiou, P.G., Black, M.P., Lammert, A.C., Baucom, B.R., Narayanan, S.S. 2011. "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.

Ghosh, P.K., Tsiartas, A., Narayanan, S.S., 2010. Robust voice activity detection using long-term signal variability. IEEE Trans. Audio, Speech, Lang. Process. 19, 600–613.

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

Gobl, C., Chasaide, A.N., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Comm. 40, 189–212.

Gonzaga, G.C., Campos, B., Bradbury, T., 2007. Similarity, convergence, and relationship satisfaction in dating and married couples. J.Personal. Soc. Psychol. 93, 34–48.

Gottman, J.M., Krokoff, L.J., 1989. Marital interaction and satisfaction: A longitudinal view. J. Consult. Clin. Psychol. 57, 47–52.

Gottman, J.M., Markman, H., Notarius, C., 1977. The topography of marital conflict: A sequential analysis of verbal and nonverbal behavior. J. Marriage Fam. 39, 461–477.

Grimm, M., Kroschel, K., Mower, E., Narayanan, S.S., 2007. Primitives-based evaluation and estimation of emotions in speech. Speech Comm. 49, 787–800.

Han, K.J., Kim, S., Narayanan, S.S., 2008. Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization. IEEE Trans. Audio, Speech, Lang. Process. 16, 1590–1601.

Heavey, C., Gill, D., Christensen, A. 2002. Couples interaction rating system 2 (CIRS2). University of California, Los Angeles. Los Angeles, CA, USA.

Heyman, R., 2001. Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. Psychol. Assess. 13, 5–35.

Hops, H., Wills, T.A., Patterson, G.R., Weiss, R.L. 1971. Marital Interaction Coding System. Technical Report. University of Oregon. Eugene, Oregon, USA.

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

Jones, J., Christensen, A. 1998. Couples interaction study: Social support interaction rating system. University of California, Los Angeles. Los Angeles, CA, USA.

Jurafsky, D., Ranganath, R., McFarland, D. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In: Human Language Technologies, Boulder, CO, USA. pp. 638–646.

Juslin, P.N., Scherer, K.R., 2005. Vocal Expression of Affect. In: The New Handbook of Methods in Nonverbal Behavior Research. Oxford University Press, Oxford, UK, chapter 3, pp. 65–135.

Karney, B.R., Bradbury, T.N., 1995. The longitudinal course of marital quality and stability: A review of theory, methods, and research. Psychol. Bull. 118, 3–34.

Katsamanis, A., Black, M.P., Georgiou, P.G., Goldstein, L., Narayanan, S.S. 2011a. SailAlign: Robust long speech-text alignment. In: Very-Large-Scale Phonetics Workshop, Philadelphia, PA, USA.

Katsamanis, A., Gibson, J., Black, M.P., Narayanan, S.S. 2011b. Multiple instance learning for classification of human behavior observations. In: Affective Computing and Intelligent Interaction, Memphis, TN, USA.

Keen, D., 2005. The use of non-verbal repair strategies by children with autism. Res. Dev. Disabil. 26, 243–254.

Kerig, P.K., Baucom, D.H. (Eds.), 2004. Couple Observational Coding Systems. Lawrence Erlbaum, Mahwah, NJ, USA.

Lee, C.M., Narayanan, S.S., 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. 13, 293–303.

Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S.S. 2009. Emotion recognition using a hierarchical binary decision tree approach, in: Proc. Interspeech, Brighton, UK. pp. 320–323.

Lee, C.C., Black, M.P., Katsamanis, A., Lammert, A.C., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.S., 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: Proc. Interspeech, Makuhari, Japan. pp. 793–796.

Lee, C.C., Katsamanis, A., Black, M.P., Baucom, B.R., Georgiou, P.G., Narayanan, S.S. 2011. An analysis of PCA-based vocal entrainment measures in married couples' affective spoken interactions. In: Proc. Interspeech, Florence, Italy.

Margolin, G., Oliver, P.H., Gordis, E.B., O'Hearn, H.G., Medina, A.M., Ghosh, C.M., Morland, L., 1998. The nuts and bolts of behavioral observation of marital and family interaction. Clin. Child Fam. Psychol. Rev. 1, 195–213.

Margolin, G., Gordis, E.B., Oliver, P.H., 2004. Links between marital and parent–child interactions: Moderating role of husband-to-wife aggression. Dev. Psychopathol. 16, 753–771.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157.

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

Moreno, P.J., Joerg, C., van Thong, J.M., Glickman, O. 1998. A recursive algorithm for the forced alignment of very long audio segments. In: Proc. ICSLP, Sydney, Australia.

Murray, K. 2001. A study of automatic pitch tracker doubling/halving "errors". In: SIGdial Workshop on Discourse and Dialogue, Aalborg, Denmark.

O'Brien, M., John, R.S., Margolin, G., Erel, O., 1994. Reliability and diagnostic efficacy of parents' reports regarding children's exposure to marital aggression. Violence and Victims 9, 45–62.

Ranganath, R., Jurafsky, D., McFarland, D. 2009. It's not you, it's me: Detecting flirting and its misperception in speed-dates. In: Conference on Empirical Methods in Natural Language Processing, Suntec City, Singapore. pp. 334–342.

Rozgić, V., Xiao, B., Katsamanis, A., Baucom, B., Georgiou, P.G., Narayanan, S.S., 2010. A new multichannel multimodal dyadic interaction database. In: Proc. Interspeech, Makuhari, Japan. pp. 1982–1985.

Rozgić, V., Xiao, B., Katsamanis, A., Baucom, B., Georgiou, P.G., Narayanan, S.S. 2011. 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. pp. 2368–2371.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L. 2007. The relevance of feature type for automatic classification of emotional user states: Low level descriptors and functionals. In: Proc. Interspeech, Antwerp, Belgium. pp. 2253–2256.

Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., Arsic, D., Rigoll, G. 2008. 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. pp. 4501–4504.

Schuller, B., Steidl, S., Batliner, A. 2009a. The Interspeech 2009 emotion challenge. In: Proc. Interspeech, Brighton, UK. pp. 312–315.

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

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.S. 2010. The Interspeech 2010 paralinguistic challenge. In: Proc. Interspeech, Makuhari, Japan. pp. 2794–2797.

Sevier, M., Simpson, L.E., Christensen, A. 2004. Demand/withdraw interaction coding. Lawrence Erlbaum, Mahwah, NJ, USA. Couple observational coding systems, pp. 159–172.

Sevier, M., Eldridge, K., Jones, J., Doss, B.D., Christensen, A., 2008. Observed communication and associations with satisfaction during traditional and integrative behavioral couple therapy. Behavior Ther. 39, 137–150.

Shoham, V., Rohrbaugh, M., Stickle, T., Jacob, T., 1998. Demand-withdraw couple interaction moderates retention in cognitive-behavioral versus family-systems treatments for alcoholism. J. Fam. Psychol. 12, 557–577.

Tranter, S.E., Reynolds, D.A., 2006. An overview of automatic speaker diarization systems. IEEE Trans. Audio, Speech, Lang. Process. 14, 1557–1565.

Traunmüller, H., Eriksson, A. 1994. The frequency range of the voice fundamental in the speech of male and female adults. Technical Report. Linguistics Department, University of Stockholm. Stockholm, Sweden.

Vinciarelli, A., Pantic, M., Bourlard, H., 2009. Social signal processing: Survey of an emerging domain. Image Vision Comput. 27, 1743–1759.

Williams-Baucom, K., Atkins, D., Sevier, M., Eldridge, K., Christensen, A., 2010. "You" and "I" need to talk about "us": Linguistic patterns in marital interactions. Pers. Relationship. 17, 41–56.

Yildirim, S., Narayanan, S.S., Potamianos, A., 2010. Detecting emotional state of a child in a conversational computer game. Comput. Speech Lang. 25, 29–44.