"That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?

Panayiotis G. Georgiou[1], Matthew P. Black[1], Adam C. Lammert[1], Brian R. Baucom[2], and Shrikanth S. Narayanan[1,2]

[1] Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA
[2] Department of Psychology, Univ. of Southern California, Los Angeles, CA, USA
http://sail.usc.edu

Abstract. Psychology is often grounded in observational studies of human interaction behavior, and hence on human perception and judgment. There are many practical and theoretical challenges in observational practice. Technology holds the promise of mitigating some of these difficulties by assisting in the evaluation of higher level human behavior. In this work we attempt to address two questions: (1) Does the lexical channel contain the necessary information towards such an evaluation; and if yes (2) Can such information be captured by a noisy automated transcription process. We utilize a large corpus of couple interaction data, collected in the context of a longitudinal study of couple therapy. In the original study, each spouse was manually evaluated with several session-level behavioral codes (e.g., level of acceptance toward other spouse). Our results will show that both of our research questions can be answered positively and encourage future research into such assistive observational technologies.

Keywords: psychology, human behavior, observational studies, lexical features, categorization, couple therapy, behavioral signal processing, behavioral informatics, BSP

1 Introduction

Human perceptual judgments form the basis for many kinds of psychological evaluations. Social therapies rely on a methodology involving careful observation and assessment of social, affective, and communicative behavior. While some of these judgments can be made in real-time during the interaction, oftentimes the interaction is recorded for offline hand coding of relevant observational events, especially for training purposes and research. In family studies research and practice, psychologists rely on a variety of established coding standards [8]. There are many examples of standardized coding schemes [4], all with the aim of producing accurate, consistent ratings of human behavior by human annotators.

This manual coding is a costly and time consuming process. First, a detailed coding manual must be created, which often requires several design iterations. Then, multiple coders, each of whom has his/her own biases and limitations, must be trained in a consistent manner. The process is mentally straining and the resulting human agreement is often quite low [8].

Given all of the practical challenges, there is clearly a strong motivation for finding technological solutions to manual coding. The ability to automatically estimate perceptual judgments and to predict the relevant behavioral codes could provide huge savings. This is one goal of human behavioral signal processing, which uses technology to extract human-centered information, including affect and emotions [6, 10].

In terms of designing a system for automatic coding, there are many possible signals to consider. These include video (e.g., gesture, body language), audio (e.g., acoustic properties of speech and other vocal cues), and transcripts (e.g., lexical features). Our working corpus – described below – offers us with access to all of these signals. It was collected in the context of a longitudinal study of couple therapy, where husband-wife pairs participated in spontaneous discussions about pertinent marital issues. The interaction sessions had been rated with a variety of high-level behavioral codes, including blame, acceptance, global negative and positive affect, as well as humor and sadness.

Recently published work from our group utilized audio features from this corpus to predict these high-level behavioral codes [1, 7]. In this paper, we investigate lexical features. We expect that lexical features contain rich information about these codes, and more broadly about the overall interactions. Towards establishing that we will pursue two goals:

1. to demonstrate that the specific behavioral codes sought by psychologists in couple interactions are expressed strongly through the lexical channel, and hence, lexical classification will be useful in this domain.
2. to show that despite the highly reverberant, noisy, spontaneous, emotional, and disfluent nature of the interactions, the automated lexical classification process from the audio signals can retain sufficient information towards extracting the behavioral codes.

Transcripts have proved useful in analyzing human social interactions [9]. Indeed, when coding manuals are used to help standardize a rating procedure, they are often very specific about the kind of phrasing and word choices, which are indicative of a particular behavioral code. The manuals used for generating our current corpus (see Section 2) are good examples. For instance, [3] states that "explicit blaming statements," such as "you made me do it," should warrant high ratings for the blame code. However it is unclear how context dependent is human interpretation of these transcripts, and hence how useful are shorter lexical units, so our first research goal will explore this question.

Obviously, automatic transcription is desirable since manually producing transcripts is laborious, albeit a very small fraction of the coding effort. Since the ultimate goal is not to produce a faithful reference transcript but to estimate behavioral codes, canonical automatic speech recognition (ASR) performance

(e.g. word error rate) is not a suitable evaluation metric. We propose to proceed with our second research goal through probabilistic representations of ASR hypotheses but without a single reference transcript.

In Section 2 of this paper, we describe the couples' interaction corpus. Section 3 presents our methodology and results. In Section 4, we present discussion and future directions.

2 The Corpus

The corpus consists of audio and video, recorded during sessions of real couples interacting. The recordings were made in conjunction with a longitudinal study at the University of California, Los Angeles and at the University of Washington [2]. For the study, over 130 husband-wife pairs were recruited to receive couple therapy for a period of one year. Each couple was recorded three separate times: before therapy began, 26 weeks into therapy and two years after therapy had finished. During each session, the couples discussed a problem in their relationship. The couple spent ten minutes on a topic of the wife's choosing, as well as ten minutes on a topic chosen by the husband. These interactions took place with the therapist out of the room. The couples were married an average of 10.0 years (SD = 7.7), with the age of participants ranging from 22 to 72 years old. The median age for men was 43 years (SD = 8.8), and 42 years (SD = 8.7) for women. Participants were college-educated, on average, with a median of 17 years of education (SD = 8.7). The sample was 77% Caucasian, 8% African American, 5% Asian or Pacific Islander, 5% Latino/Latina, 1% Native American, and 4% Other [2].

Transcriptions were made of the audio recordings for each interaction that constituted the data used for this study. The speaker for each turn (husband or wife) was explicitly labeled. Efforts were made to keep the transcriptions as faithful to the audio as possible. An example fragment from one transcript can be seen in Table 1. It should be noted that nonverbal communication (e.g., laughing, throat clearing) was transcribed. However, names and proper nouns were de-identified for the sake of privacy. Unintelligible regions were also marked by the transcribers. Only 0.98 percent of the words were de-identified or marked unintelligible. In portions with overlapping speech, transcribers attempted to separate out words from each speaker. These portions were not explicitly marked.

For each session, both spouses were evaluated according to 33 codes, designed to rate an individual's interaction. The Social Support Interaction Rating System (SSIRS) measures both the emotional component of the interaction, as well as the topic of conversation. Its 20 codes are broken into four categories: affectivity, dominance/submission, features of the interaction, and topic definition [4]. The 13 codes in the Couples Interaction Rating System 2 (CIRS2) were specifically designed to capture perception relevant for conversations involving a problem in the relationship [3]. Three to four evaluators made their judgments after watching the video recording. All evaluators underwent training, in an effort to standardize the coding process. Ratings were expressed on an integer scale

Table 1. *An example fragment from one transcript. In this particular interaction, the wife received a low rating for acceptance and a high rating for blame and negativity, and the husband received a low rating for humor and a high rating for positivity and sadness.*

Partner	Transcript
H	WHAT DID I TELL YOU YOU CAN DO THAT AH AND EVERYTHING
W	BUT WHY DID YOU ASK THEN WHY DID TO ASK
H	AND DO IT MORE AND GET US INTO TROUBLE
W	YEAH WHY DID YOU ASK SEE MY QUESTION IS
H	MM HMMM
W	IF IF YOU TOLD ME THIS AND I AGREE I WOULD KEEP TRACK OF IT AND EVERYTHING
H	THAT'S THAT'S
W	**THAT'S AGGRAVATING VERY AGGRAVATING**
H	A BAD HABIT THAT
W	VERY AGGRAVATING
H	CAUSES YOU TO THINK THAT I DON'T TRUST YOU
W	THAT'S EXACTLY WHY THAT'S ABSOLUTELY THE WAY IT IS
H	AND IF I DON'T THE REASON FOR THAT IS AH
W	I DON'T CARE THE REASON YOU GET IT I GET IT TOO
H	THE REASON IS THE LONG TERM BAD PERFORMANCE
W	YEAH AND YOU KNOW WHY
H	MM HMMM
W	ALL YOU GET IS A NEGATIVE REACTION FROM ME

from 1 to 9. Here, we analyze a subset of codes which represent contrasting pairs with high inter-evaluator agreement. Codes which contrast conceptually do not necessarily contrast in rating, however. In particular, it is possible for an individual to receive similar scores on contrasting codes if each is displayed in a salient way. The codes considered in the paper include: level of *acceptance* and *blame* toward the other spouse, global *positive* and *negative*, as well as level of *sadness* and *humor*.

Further, for this work we chose to formulate the problem as a binary classification task, in which we wanted to automatically identify the two extremes of a particular code (e.g., high and low blame). To that end, we gathered 280 sessions, corresponding to the 70 highest and lowest ratings for both husband and wife. Each code was considered separately in this regard, resulting in six binary classification tasks. For all experiments in this study, we trained gender-independent models.

3 Methodology and Results

In this section we will present the mathematical methodology and results for addressing the two questions above. We start by providing the maximum likelihood classifier formulation and results on reference transcripts. Then we describe our available audio and the classification method on the automatically derived lexical features and results.

Our goal is to show the usefulness of lexical features in accurate prediction of the high-level behavioral ratings. For this first study we chose a simple classifier with unigram features. We expect that the frequencies of lexical terms (e.g., n-gram features) are crucially informative. As a first study we only deal with unigram features to minimize data sparsity issues that appear with higher order

n-grams. More specifically, we make use of only the unigram frequencies of the individual being rated, while ignoring those of their spouse. Since information about one of the partners behaviors can arguably be highly dependent on the behavior of the other partner we intend to study that in future work.

3.1 Maximum Likelihood Classifier

In a maximum likelihood framework binary classification task we want to select the code that maximizes

$$p(\text{Code 0 or 1}|\text{Transcript}) = p(C_0 \text{ or } C_1|T) \tag{1}$$

Alternatively for $i = \{0, 1\}$:

$$C_i = \arg\max_{Ci} \frac{p(T|C_i)p(C_i)}{P(T)} \tag{2}$$

$$= \arg\max_{Ci} p(T|C_i)p(C_i) \tag{3}$$

For the purposes of this work as we described above we chose a balanced data set so $p(C_0) = p(C_1)$. Therefore the decision can be re-written as:

$$\frac{p(T|C_0)}{p(T|C_1)} = \begin{cases} > 1 \Leftrightarrow C_0 \text{ true} \\ < 1 \Leftrightarrow C_1 \text{ true} \\ = 1 \Leftrightarrow \text{no decision} \end{cases} \tag{4}$$

Given the use of unigrams

$$p(T|C_i) = \prod_{\forall w_j \in T} p(w_j|C_i) \tag{5}$$

where w_j is the j^{th} word in the transcript. As we can see because of the product term, this estimator is very sensitive to data sparsity. For instance if we have no observations in the training data of w_j then $p(T|C_i)$ will be zero. To address this we use the commonly used technique of smoothing with statistics derived from generic data, often called *Universal Background Model* (UBM), here denoted by B:

$$p(T|C_i) = \prod_{\forall w_j \in T} \left[(1 - \lambda)\, p(w_j|C_i) + \lambda p(w_j|B)\right] \tag{6}$$

The background model also serves to boost the importance of lexically salient regions. As $\lambda \to 1$ only the words with significantly different probabilities within the two domains – and arguably more important – will contribute to the decision.

3.2 Classification on Reference Transcripts

For seeking our first research goal of establishing the usefulness of lexical information for behavioral code classification, we follow the process outlined in

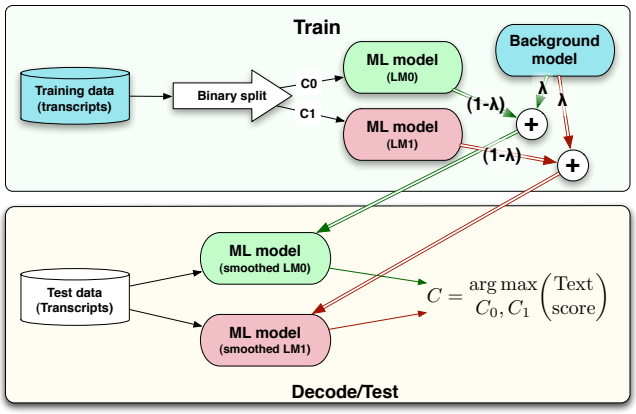

Fig. 1. *Overview of the classification process from manually generated transcripts of the interactions.*

Fig. 1. Through a leave-one-couple-out process we train a maximum likelihood model of unigram probabilities for the two classes, and smooth it with a UBM. Each of these models can be used to score the test transcript.

As mentioned above we chose to work with the 280 sessions that had received the extreme behavior codes. Given that this sometimes may include the same couple twice this results in about 85 unique couples per code. For instance in *blame* the 100 sessions resulted in 89 couples. Model selection was done with leave-one-couple-out cross-validation, rather than leave-one-transcript-out, to avoid the possibility of some speakers appearing in both the training and test sets (e.g. *blame* resulted in 89 folds). For comparison purposes, we calculated the percentage correctly classified.

Table 2 shows the performance for the six codes. From the experimental results, it can be seen that regardless of λ lexical information support behavioral code prediction.

Separability by Human experts and Machine As can be seen by Table 2, the codes Humor and Sadness perform the worst amongst the six codes chosen. Fig. 2 provides insight as to why. As we can see, the human annotators had the least discrimination in those two codes with the positive (right) part of the distribution exhibiting a large spread. In fact one could argue that by looking at the data that Humor for instance did not exhibit a bi-modal distribution in the original annotations and that our choice of the top/bottom highest ratings

Table 2. *Results of classification using reference transcripts for different λ.*

Results on reference transcript (% correct)											
code vs λ	**0.01**	**0.05**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.95**	**0.99**
acceptance	91.4	91.0	91.0	90.0	90.3	89.2	88.5	87.5	86.4	75.3	60.5
blame	91.0	91.4	91.8	91.0	90.3	89.2	89.2	88.5	88.2	78.1	63.4
humor	71.3	72.4	72.0	71.3	69.5	69.9	67.5	67.0	65.2	61.6	57.3
negative	83.8	84.9	86.7	86.7	86.4	85.7	86.0	86.0	85.3	74.9	60.2
positive	89.6	89.6	89.6	88.9	87.5	87.8	87.8	87.5	87.8	76.7	63.8
sadness	59.0	61.6	60.9	61.3	60.6	60.2	58.8	59.5	59.1	57.7	58.5

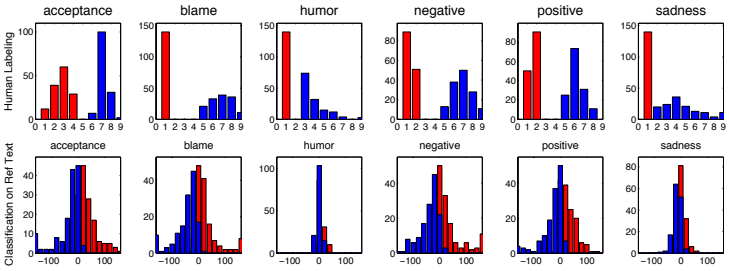

Fig. 2. *Distribution of data based on (top) the average ratings provided by multiple human experts and (bottom) the difference in log-likelihoods of the ML model for* $\lambda = 0.5$*. As can be seen codes where annotators had minimal separation also result in the greatest overlap by the ML model.*

may have been – for these two codes – a necessary but not necessarily the best choice.

3.3 Audio Segmentation

Prior to decoding the speech into words we need to separate the audio of the two participants. In this paper and in our previous work [1, 7], we decided to exploit the available transcriptions towards that task through forced alignment. The process employs *SailAlign* [5] that can be summarized as a recursive speech recognition and alignment of the ASR-output with reference transcript.

Although for our previous work this process offered the advantage of higher accuracy of word boundaries, it also has a higher rejection rate as words may be marked as un-aligned. Many times this can occur even if the word is in the middle of a continuous speech segment by the same speaker. Hence not all the spoken audio is used in the automated classification process. Of the original 569 sessions, 372 met both the threshold of 5 dB SNR and >55% aligned audio, which left us 62.8 hours of data across 104 unique couples. This reduction in data resulted in about 70% less data for the classification of section 3.4 compared to the transcript analysis in section 3.2.

This is clearly going to have some effect on the observed performance drop in the case of classification directly from audio. Note that all the transcripts are used in the classification from reference transcripts in order to set the upper bound of possible performance.

3.4 Classification on Audio Signal

For pursuing our second research goal of behavioral code classification from noisy automatically obtained speech transcriptions we follow the process outlined in Fig. 3. As before we train our models with smoothing from a UBM. In this case however we also create a class-independent ML model (LM_{ASR}) that can be used for the automatic speech recognition process. Note that the specific dataset has a very wide range of acoustic characteristics and as such our ASR

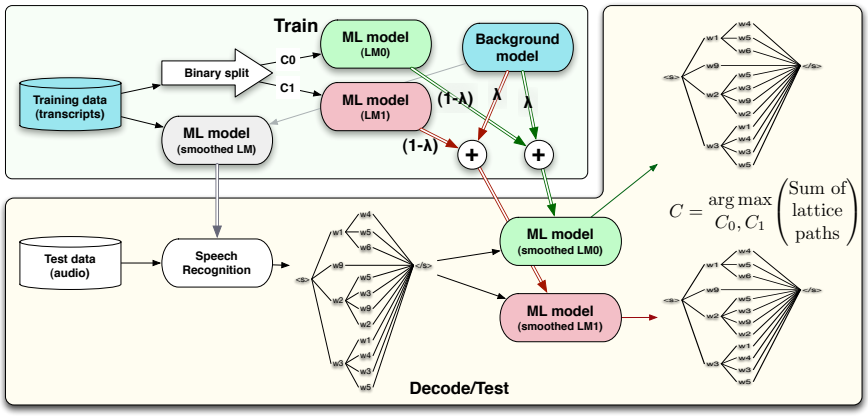

Fig. 3. *Overview of the classification process without human transcripts through the use of ASR lattices.*

was not optimized for the domain, and online adaptation was switched off. We used OtoSense for our ASR (a SAIL implementation) and acoustic models based on WSJ and HUB4. LM_{ASR} similarly was not optimized in any fashion, except in including the training data and a background model, as it is only used for lattice pruning. The word error rate (WER) varied extremely with different sessions with most lying between a WER of 40-90%.

Our assumption towards establishing the procedure for implementing our second research goal is that the noise introduced through the ASR process is independent from the couple behaviors. We believe this assumption to be a valid one given that the acoustic mismatch includes reverberation, environment and sensing noise, and speaker-specific acoustic pattern mismatch.

Therefore at the test phase we decode and produce a lattice using the same ASR (acoustic models and LM_{ASR}) and then replace the language model scores in the lattice with the class LM values. The final step is to decode the two resulting lattices and find the score of the N-best paths. For the purposes of this paper we used N=100 (an unoptimized parameter).

Table 3 shows the performance for the six codes and for different values of λ. As we can see there is a significant degradation relative to the values in Table 2, however we can also note that for most codes the performance is significantly

Table 3. *Results of classification using lexical analysis of audio for different λ.*

Results through ASR lattices (% correct)											
code vs λ	**0.01**	**0.05**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.95**	**0.99**
acceptance	71.4	72.9	75.4	73.6	73.6	73.2	71.8	71.1	68.9	64.6	63.6
blame	75.0	76.8	77.9	78.6	78.2	77.9	76.8	76.4	73.9	67.5	63.9
humor	57.9	58.6	58.6	57.5	57.1	56.4	57.9	56.1	55.4	55.0	50.7
negative	64.3	66.1	69.6	71.1	70.4	69.3	69.3	67.9	65.7	60.7	58.9
positive	72.9	73.2	74.6	74.6	72.5	72.9	73.9	73.6	71.4	66.1	64.6
sadness	52.5	55.0	55.7	52.1	50.4	50.7	51.1	51.8	52.1	54.3	52.1

Table 4. *The unigrams with most impact towards the correct classification of blame for one of the cross-validation folds.*

Most blaming words in terms of discriminative contribution				Least blaming words in terms of discriminative contribution			
Word	No Bl. log prob	Blame log prob	Δ	Word	No Bl. log prob	Blame log prob	Δ
YOU	-95.49	-85.88	-9.61	EXPECTS	-16.70	-17.84	1.14
YOUR	-51.24	-47.18	-4.06	CONSIDERATION	-16.11	-17.31	1.21
ME	-40.27	-37.74	-2.53	KNOW	-35.10	-36.62	1.53
TELL	-33.97	-32.46	-1.51	INABILITY	-16.76	-18.32	1.55
ACCEPT	-25.44	-23.99	-1.45	SESSION	-20.51	-22.07	1.56
CARING	-27.05	-25.91	-1.14	OF	-44.50	-46.26	1.76
KITCHEN	-21.22	-20.21	-1.02	ANTICIPATION	-22.22	-24.21	2.00
TOLD	-29.04	-28.19	-0.85	THINK	-35.70	-37.77	2.07
NOT	-40.32	-39.59	-0.73	WE	-29.39	-31.75	2.36
WHAT	-51.47	-50.77	-0.69	I	-99.92	-102.49	2.57
INTIMACY	-43.16	-42.53	-0.63	THAT	-91.30	-93.97	2.67
IT	-42.70	-42.18	-0.52	UM	-64.75	-70.76	6.01

better than chance (50%). Regardless of λ it is clear that Hypothesis 2 holds, even with a generic ASR and even with the data loss due to automated segmentation.

3.5 Lexical Significance

In a parallel analysis and in collaboration with our psychologist partners we also looked into whether specific words offered insights into specific behavioral codes. For instance Table 4 shows that specific words can carry a lot of insight towards the behavioral codes. As we can see the word YOU, which appeared 59 times, had the most contribution towards the blame decision, while the word UM (23 times) scored as the least blaming unigram ($\lambda = 0.4$). The mathematical analysis enables us to easily identify important terms that we will follow up with detailed experimental and psychological inquiry.

4 Discussion

Our goal in this work was to establish the usefulness of lexical features for the purpose of machine classification of human behavior in couple interactions. The answers to our first question clearly show that even a simple ML unigram based classifier can achieve good classification accuracy. The experiments related to our second research question established that despite the very large WER of the ASR, significant behavior information is contained in the noisy lattices.

In sum these experiments show that lexical information is an important information stream for the important task of automatic behavioral coding in couple interactions. There are a number of improvements toward improving automated classification that we plan on pursuing in our follow-up work.

First we want to address the relative importance of information within a specific data stream through salience detection. Currently as mentioned above lexically salient words are given more importance, however salience through acoustic information is not considered.

In addition we want to consider higher order n-gram streams. These can not be used directly with our current corpus due to data sparsity, so we plan to

pursue techniques to address that including data mining for richer models, appropriate smoothing techniques (e.g. Kneser-Ney), and the fusion of the decision of different lexical classifiers (e.g. along salience sensitivity and n-gram order).

At the system level we want to investigate alternative classifiers. We already have initial results using a SVM classifier, but without using a UBM and noted that it under-performs the technique reported here. We intend to combine the smoothing with a SVM classifier in our future work.

Finally in our previous work [1] we presented an acoustics based classification framework. The fusion of the two information streams can potentially provide great benefits. In addition fusion can take place at various temporal resolutions from word level fusion to session level.

5 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 sharing the data.

References

1. Black, M., Katsamanis, A., Lee, C.C., Lammert, A., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.: Automatic classification of married couples' behavior using audio features. In: Proc. Int'l Conf. on Speech Communication and Technology (2010)
2. Christensen, A., Atkins, D., Berns, S., Wheeler, J., Baucom, D., Simpson, L.: Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. Journal of Consulting and Clinical Psychology 72(2), 176–191 (2004)
3. Heavey, C., Gill, D., Christensen, A.: Couples interaction rating system 2 (CIRS2). University of California, Los Angeles (2002), http://christensenresearch.psych.ucla.edu/
4. Jones, J., Christensen, A.: Couples interaction study: Social support interaction rating system. University of California, Los Angeles (1998), http://christensenresearch.psych.ucla.edu/
5. Katsamanis, A., Black, M.P., Georgiou, P.G., Goldstein, L., Narayanan, S.S.: SailAlign: Robust long speech-text alignment. In: Very-Large-Scale Phonetics Workshop (Jan 2011)
6. Lee, C., Narayanan, S.: Towards detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13(2), 293–302 (March 2005)
7. Lee, C.C., Black, M., Katsamanis, A., , Lammert, A., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.: Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: Proc. Int'l Conf. on Speech Communication and Technology (2010)
8. Margolin, G., Oliver, P., Gordis, E., O'Hearn, H., Medina, A., Ghosh, C., Morland, L.: The nuts and bolts of behavioral observation of marital and family interaction. Clinical Child and Family Psychology Review 1(4), 195–213 (1998)
9. Ranganath, R., Jurafsky, D., McFarland, D.: It's not you, it's me: Detecting flirting and its misperception in speed-dates. In: EMNLP (2009)
10. Yildirim, S., Narayanan, S., Potamianos, A.: Detecting emotional state of a child in a conversational computer game. Computer Speech & Language (2010)