A Dynamic Model for Behavioral Analysis of Couple Interactions using Acoustic Features

Wei Xia[1], *James Gibson*[1], *Bo Xiao*[1], *Brian Baucom*[2], *Panayiotis Georgiou*[1]

[1]University of Southern California, Los Angeles, CA, USA
[2]The University of Utah, Department of Psychology, UT, USA

{weixia,jjgibson,boxiao}@usc.edu brian.baucom@psych.utah.edu georgiou@sipi.usc.edu

Abstract

Observational therapy is an important element of mental health that relies on a detailed assessment of multiple behavioral cues. Behavioral coding for research in the field is unfortunately often at session-level resolution due to the inherent cost of labeling and human subjectivity. Being able to model the interlocutors' behavior at a fine temporal resolution and analyze the effect of such behavioral changes in the gestalt perception can help psychologists better understand the behavioral mechanism. In this paper, we propose a method to model the dynamically evolving behavior of interlocutors during couple interactions. We firstly present a static behavioral model based on the local decisions with global fusion, and investigate the impact of the frame length to provide effective global evaluations. We then propose a two-layer sequential Hidden Markov Model to capture local state transitions. We use the corpus of Couple Therapy interactions as a case study, finding that an interlocutor does not express a single behavior throughout a conversation, and there are temporal correlations between neighboring frames. We show that dynamic models can achieve up to 10% relative improvement, compared to static models. This suggests that the human behavioral interaction is a non-linear process, and the resulting latent-state labels may provide new insights to domain experts.
Index Terms: Behavioral Signal Processing, Couple Therapy, Hidden Markov Model, Sequential Learning

1. Introduction

Modeling and understanding couples behavior is a challenging issue that is central to the Couple Therapy research which is demonstrated to be an effective way to improve marital relationships. This is a complicated learning problem that requires a great deal of manual annotations and domain insights. One fundamental task of couple therapy is to observe, identify, and analyze behaviors during interactions so that psychologists can provide effective and specific treatment.

During the last decade, we have seen a range of developments towards assessing the human state. Pantic et al. [1, 2] provided a survey of automatic analysis of social signals using both verbal and non-verbal cues. They categorize information conveyed by humans into five types: affective states, manipulators, emblems, illustrators, and regulators. Schuller et al. [3, 4, 5] provided detailed analysis about the feature engineering for automatic emotion recognition. They proposed various low-level descriptors (LLDs) and hierarchical functionals of acoustic and linguistic features and applied different feature selection and dimension reduction techniques to identify the optimal subset achieving the best classification performance. The work in social signal processing and affect recognition, however, has not used domain-specific analysis of the human behavior.

Recently, the new area of "Behavioral Signal Processing" [6, 7, 8] has been established, which focuses on gathering, analyzing and modeling multi-modal behavioral signals based on lexical [7], speech [9], and visual [10] signal processing and machine learning techniques to aid experts in the psychology domain. The challenges are many-fold: Human behavioral coding is based on pre-defined behavioral dimensions, and result from human annotator's subjective perceptions of short and often sparse observation windows. Further, annotations are most often at very coarse scales and integrate human expressions, with their many local variations and transitions, in subjective, complex and non-linear ways.

For example in the couple therapy domain [11], Black et al. [9, 12] proposed an automatic behavioral coding scheme for couples' interactions using acoustic features. They employed global prosodic and spectral features to predict the behaviors of each spouse across six dimensions for the whole 10-minute interaction and achieved good performance. Their proposed model, however, does not provide any finer resolution information such as where these behaviors occur within the 10-minute interaction. Such information is vital in informing experts for couple-specific behavioral patterns. Katsamanis et al. [13] demonstrated that some salient segments of the couple interaction can be more informative than others in terms of behavior, and they used a multiple instance learning [14] framework to exploit those salient areas to improve behavioral classifications.

The above models do not provide fine resolution information or take advantage of human dynamics. Early work into this is in the lexical domain, *e.g.,* Chakravarthula et al. [15] proposed a language based dynamic model to investigate the undergoing local changes in the behavioral state of a spouse within a conversation. Learning a dynamic model from acoustic data that can exploit behavioral state transitions may further contribute to modeling dyadic interactions and is the goal of our work.

2. Overview

In psychology, annotations are often gestalt subjective ratings of the whole interaction. For example, an expert can rate an interlocutor as having "Negative Affect" for a 10-minute interaction; similarly a naive observer may say that the interlocutor is "rude". This has been approximated in most existing work as a generating function that assumes all data generated within the session is of a single class (say C1 ="negative"). This is shown in Fig. 1-left.

The dynamic evolution process of human behavior through the interaction is depicted in Fig. 1-right. We describe the local behavioral states as a first order stationary Markov process and

This work is supported by NSF and DoD.

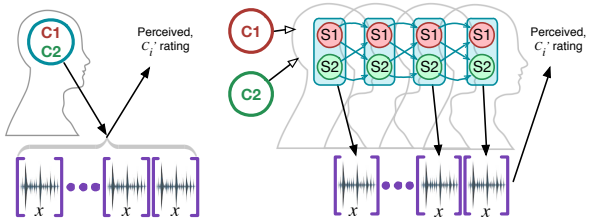

Figure 1: *Overview of the proposed behavioral modeling framework. Left: Static Behavioral Model; Right: Dynamic Behavioral Model. (C: true global state, C_i': perceived rating)*

assume the existence of temporal correlations between neighboring frames. As shown in the figure, previous hidden states can propagate to the current state and the behavioral observations are generated by the hidden local states. What is perceived to be a certain behavior is a constantly evolving process that may move from positive (say S2 = "how are you today") to less positive (say S1 = "I don't care"). The overall perception C_i' can be a highly non-causal and a non-linear integration of the observations of such states. This proposed process can be better modeled through a dynamic system. An example would be a Hidden Markov Model (HMM) that models the transitions over time and the underlying observations are established by the local classifiers. This is precisely our contribution.

In Section 3.1 we establish the *Static Behavioral Model* (SBM) as the baseline model. We employ three machine learning algorithms to derive local decisions and fuse those to obtain a single session-level decision to evaluate against the human annotation C_i'. We proceed in Section 3.2 to describe our proposed implementation of the *Dynamic Behavioral Model* (DBM) illustrated in Fig. 1-right. To validate our work, we employ the couple therapy corpus as a case study in Section 4, and we provide detailed explanations of our experiments in Section 5 and corresponding results and discussions in Section 6. We show that the dynamic behavioral models outperform the static behavioral models by a large margin and provide informative latent-state labels. Finally, we conclude in Section 7 with future work.

3. Behavioral Modeling

3.1. Static Behavioral Models

We describe the behavioral model training and evaluation in this section. In static behavioral models (SBM), the underlying assumption is that the interlocutor consistently generates the same behavior throughout the interaction. This means that if we analyze any short-term window, we can capture the same behavior as if we analyze any other window, and that behavior is very consistent with the perceived, annotated behavioral rating. This allows us to train local classifiers independently (Sec. 3.1.1) and express the session-level behavior as a combination (Sec. 3.1.2) of the local discrimination results.

3.1.1. Local Behavioral Descriptors

We compare three algorithms for classifying behavior from each short-term window: Support Vector Machine (SVM), Voted Perceptron (VP), and Fisher Linear Discrimination Analysis (FLDA). We use these three methods in order to capture both non-linear and redundant characteristics of acoustic features. SVM is popular in the emotion recognition field since it can be easily kernalized to handle non-linear data. Also, since computational static functionals of raw acoustic features have high collinearity, we use Fisher LDA to reduce potential "noisy" directions. The loss functions of SVM and Fisher LDA are given in Eqs. (1) and (2).

$$\boldsymbol{\omega}_1 = \arg\min_{\boldsymbol{\omega}} \frac{1}{2}\|\boldsymbol{\omega}\|_2^2 + C\sum_{i=1}^{m} \max\left(0, 1 - y_i \boldsymbol{\omega}^{\mathrm{T}} \mathbf{x}_i\right)^2 \quad (1)$$

$$\boldsymbol{\omega}_2 = \arg\max_{\boldsymbol{\omega}} \frac{\boldsymbol{\omega}^{\mathrm{T}} S_B \boldsymbol{\omega}}{\boldsymbol{\omega}^{\mathrm{T}} S_W \boldsymbol{\omega}} \quad (2)$$

Constant C in Eq. (1) is the regularization coefficient. S_B and S_W in Eq. (2) are the between classes scatter matrix and within classes scatter matrix respectively. Due to the natural online property of the Perceptron algorithm, we also adopt the Voted Perceptron method which is a weighted averaging ensemble of multiple perceptrons shown in Collins et al. [16]. Given a labeled training set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$, $\mathbf{x} \in \mathbb{R}^n$. we define a set of k perceptrons $\{(\boldsymbol{\omega}_1, c_1), (\boldsymbol{\omega}_2, c_2), \ldots, (\boldsymbol{\omega}_k, c_k)\}$, where $\boldsymbol{\omega} \in \mathbb{R}^n$, and c is the weight of each perceptron. If the training sample is misclassified by the kth perceptron, not only do we update the coefficient $\boldsymbol{\omega}_k$, but also the weight c_k. Then we combine k perceptrons to predict the local behavioral state.

3.1.2. Global Behavioral Fusion

We fuse frame-level local behavioral states to get session-level global behaviors using the majority voting method. The final estimated behavioral state is the one that receives more than half of the votes.

3.2. Dynamic Behavioral Models

The *Dynamic Behavioral Model* (DBM) employs a two-level architecture: *(i)* in the first layer, we train frame level local base classifiers. *(ii)* then the predictions from the first layer are fed into a Hidden Markov Model (HMM) in the meta layer.

We assume that the behavioral observations follow a stationary distribution and the acoustic features we use for behavioral recognition are time series at successive frames, so there may exist correlations between "past" and "future" observations. During interactions, local behavioral states of interlocutors may change over time because of the propagation from the past state. For example, a person that transitions mostly through positive states, is probably perceived as positive, and his likely internal state is positive; this, however, does in no way mean that such a positive person can not have negative expressions, and that those are not important to consider. Such a person would fit a different HMM model than a negative person.

For *(ii)* we define a HMM parameter set $\lambda = (A, B, \pi)$, where

- $A = \{a_{i,j} | a_{i,j} = P(S_j|S_i)\}$ is the transition matrix, $1 \leq i, j \leq S$.
- $B = \{b_i | b_i = P(O|S_i)\}$ is the emission matrix, O can be discrete or continuous values.
- $\pi = \{\pi_i | \pi_i = P(S_i)\}$ is the initial probability array.

3.2.1. Semi-supervised Learning of DBM

In many domain corpora, due to the cost and time constraints as well as the human subjectivity, we lack fine temporal resolution annotations. We, therefore, propose an iterative *Expectation Maximization* (EM) procedure, with hard assignment, to estimate local labels from global ratings. The idea is to initialize frame-level local labels to be the same as the corresponding session-level labels and train a local classifier. We then use this

classifier in the E-step to predict local labels. In the M-step, we re-assign labels to the local frames. We then retrain our local models and repeat the E- and M- step until frame-level labels converge. We describe the stacked SVM-HMM training and decoding procedure in the following, where n is the number of couples. We replace the SVM with other local classifiers to get different dynamic models.

Algorithm 1 Stacked SVM-HMM training and decoding

for For each test-train split $i = 1 \rightarrow n$ **do**
Train:
Initialize local state labels with the session-level ratings of train set
1. Train a SVM based on local state labels
2. Predict new states of train set based on trained SVM
3. Repeat 1, 2 until predicted local labels converge
Learn HMM parameters $\lambda = \{\pi, A, B\}$ based on local labels of train set for two classes
Test:
Perform Viterbi Decoding using the two HMMs on the test data to get the likelihoods of most likely paths
If likelihood$_1 \geq$ likelihood$_0$
Predicted global label $\leftarrow 1$
end for

3.3. Joint Optimization

The previous two sections outlined a static and a dynamic approach to detect local behavioral states, where in the training stage of our DBM models, we optimize the local frame length, hyper-parameters of local classifiers, and HMM *independently* to reduce the computational complexity. To achieve the best performance, we optimize these three factors together by grid search as well as choose the best local classifier in the cross-validation step. It is also worth mentioning that the local behavioral states may not converge in the EM step if we adjust the local frame length adaptively. In this case, we set a threshold as the total difference between current and previous predicted local states. If the difference is smaller than the threshold, we stop the EM process.

4. Data Set

4.1. Couple Therapy Corpus

We use the dataset provided by the UCLA/UW Couple Therapy Research Project described in Christensen et al. [11]. This study recruited 134 distressed couples in Los Angeles, California and Seattle, Washington. All participants were required to participate in video-taped personal and relationship discussions. Three different interactions are carefully coded: before treatment, at 26 weeks and 2 years into the treatment. These 10-minute video recordings have been transcribed and annotated. In each treatment, couples participated in two discussions (10 minutes each), one relationship or personal problem selected by the husband and another by the wife. Behaviors are coded based on the Couple Interaction Rating System (CIRS) [17] and the Social Support Interaction System (SSIRS) [18]. The CIRS coding manual contains 13 codes and is mainly designed for problem-solving behaviors, and the SSIRS has 20 codes and is designed for evaluating emotional observations and supportive behaviors. Each rating system uses a scale from 1 to 9. In this work, we analyze six binarized behavioral codes with the greatest inter-annotator agreement. They are: *Acceptance, Positivity, Humor, Sadness, Blame, and Negativity*. Further details about the dataset can be found in [12].

4.2. Acoustic Feature Processing

The original data is very noisy due to the recording environment and equipment. In order to extract robust speech features, we first employ the Voice Activity Detection (VAD) described in [19] to determine whether the region is speech or noise. We only choose audio recordings with SNR $\geq$ 5dB. After pre-processing, we finally have a total of 372 sessions.

The spectral and voice quality features are extracted using OpenSMILE [20]. Pitch is extracted using the Praat software [21]. We employ rectangular windows on 10-minute audio sessions to get short-term local frames, and then compute six static functionals of the speech features to find better representations. The features and functionals are given in Table 1.

Feature Family	Feature Members
Prosody	Intensity, fundamental frequency
Spectral	MFCC [0-14], MFB [0-7]
Voice quality	Jitter, shimmer
Functionals	Min (1st percentile), max (99th percentile) range (99th percentile - 1st percentile) mean, median, standard deviation

Table 1: *Low-level acoustic features and six static computational functionals.*

5. Experimental Setup

For our experiments, we use the top and bottom 20% data of the 372 rated sessions, as we have found that, the middle 60% has lower inter-annotator agreement with respect to each behavioral code. We regard the behavioral recognition problem as a binary classification task. We assume short-term stationarity in the behavioral signal for local decisions and compare six window lengths: 2s, 10s, 20s, 30s, 40s and 50s. We employ a rectangular window with an overlap of 25% of the frame length.

We also observe that some features, such as the median and mean of f_0, have strong collinearity so we use Principle Component Analysis (PCA) to reduce the feature dimension to 120, which can explain 90% of the total variance in the dataset.

We use the implementation in libsvm [22] package to train the SVM and Kevin Murphy's HMM toolbox [23] to train the HMM. All hyper-parameters of the local classifiers are tuned using leave-one-couple-out cross-validation since each couple has multiple 10-minute conversations (2-6 sessions per couple).

6. Results and Discussion

6.1. Static Behavioral Models

We employ SVM, Fisher LDA, and Voted Perceptron with the SBM, and as we show in Table 2, Fisher LDA outperforms the other methods, albeit only by a small margin. This is likely due to the reduction of correlation stemming from the projection into a lower dimensional space, while at the same time a significant difference between the methods is difficult due to the lack of contextual information.

Further, in Fig. 2, we show the accuracy of the Static Behavioral Model, using the best classifier, *i.e.,* Fisher LDA, versus the window length. This agrees with the thin slice theory [24], which argues that brief behavioral observations can provide a large amount of information towards detecting individual's emotional states. Psychologists and annotators may

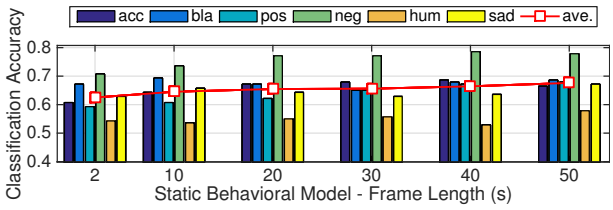

Figure 2: *Average test accuracy of the wife and husband using best SBM (FLDA) for different short-term frame lengths.*

only need to interpret a small salient part of the experimental data to provide the meaningful feedback to the couples. From the figure, we can see there is no significant difference unless we choose a very short frame length, *e.g.,* 2 seconds. By increasing the frame size, we obtain a little improvement for the classification accuracy, but we risk breaking the short-term behavioral stationarity assumption and reduce both the training frames and the temporal resolution as well.

We also notice that the significantly lower performance for *Humor* and *Sadness*, which is consistent with the existing work [7] which showed that even humans had low separation between classes for these behaviors. This suggests that individuals may not judge such implicit behaviors based on brief observations very well. On the contrary, behaviors like *Blame* are easier to identify from local sessions because people may discriminate them using salient characteristics in such behaviors.

6.2. Dynamic Behavioral Models

In Fig. 3, we show the normalized state confusion histogram from the static Fisher LDA to the dynamic Fisher LDA when the behavioral code is *Sad*. 35.71% of data classified erroneously by the SBM is still categorized mistakenly by the DBM, which is partly due to the local behavioral descriptor and the combination rule we use, but we can observe 10.71% of data classified wrongly by the SBM is corrected by the DBM. This is consistent with our assumptions that 1) the previous local state may affect the current local state. 2) the interlocutor does not express a single behavior throughout the interaction process, there may exist local variations of the behavioral state. So we build a second layer based on the HMM using the predictions of local classifiers to model the dynamic transitions over time.

The second part of Table 2 shows the performance of the Dynamic Behavior Models. This table shows the additional contextual information captured by the HMM improves the performance consistently in all cases. Dynamic Fisher LDA performs better than most other models. We take a comparison of the static SVM and the SVM-HMM approach as an example to discuss the results. We obtain an average accuracy of 68.21% for the husband, and 67.86% for the wife using the SVM-HMM

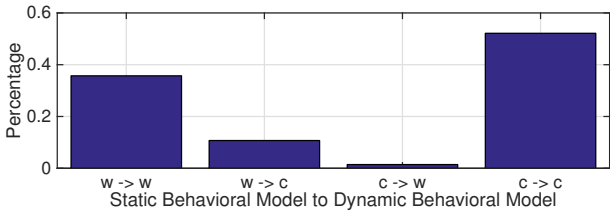

Figure 3: *State confusion percentage from SBM to DBM: w-w: wrong to wrong, w-c: wrong to correct, c-w: correct to wrong, c-c: correct to correct, when classifier = FLDA, code = sad, frame length = 50s.*

SBM	Spouse	*Acc*	*Bla*	*Pos*	*Neg*	*Hum*	*Sad*	Average
SVM	wife	64	74	65	71	54	58	64
	husband	65	64	49	78	54	61	62	63
Voted Perceptron	wife	68	75	62	71	52	61	65
	husband	66	66	60	68	60	60	63	64
Fisher LDA	wife	62	76	68	73	58	55	65
	husband	67	67	62	77	55	**64**	65	65
DBM	**Spouse**	***Acc***	***Bla***	***Pos***	***Neg***	***Hum***	***Sad***	**Average**
SVM-HMM	wife	72	73	**76**	71	57	60	68
	husband	68	**76**	66	80	61	57	68	68
VP-HMM	wife	70	76	71	72	64	**65**	70
	husband	72	71	68	75	64	54	67	69
Fisher-HMM	wife	**76**	**78**	75	**76**	**64**	62	**72**
	husband	**73**	75	**76**	**80**	**69**	59	**72**	**72**
Joint Optimization	wife	**78**	**81**	**81**	**79**	**66**	**65**	**75**
	husband	**74**	**76**	**78**	**84**	**69**	61	**74**	**75**

Table 2: *Classification accuracy (%) of static local classifiers with global fusion and dynamic sequential models based on HMM, when frame length = 20s.*

approach, improving 4.16% and 6.31% respectively compared to the static SVM method. We obtain the greatest improvement for the *Blame* behavior, indicating that this behavior might be easy to represent based on acoustic features. We get a 12.14% improvement for the husband model, suggesting there may be numerous local variations and state propagations in the *Blame* behavior. For the wife model, however, the dynamic SVM shows much less improvement, compared with the static SVM. This result indicates that females may have a greater tendency to remain in a blaming state more consistently throughout the session, while males may have more dynamic changes when expressing blame.

As expected, for *Humor* and *Sadness*, our model does not perform very well in all cases. However, it should be noted that while the performance for *Sadness* still remains low in the DBM case, the classification accuracy for *Humor* using DBM improves by a very large margin. For instance, the accuracy for *Humor* with static FLDA is at an average of 56.43% (57.86% and 55.00% for the wife and husband respectively), and it jumps to 66.43% with FLDA-HMM (64.29% and 68.57%), showing a 10% absolute improvement. This suggests that *Humor* is highly contextual. People are less likely to express humor constantly, while they are likely to remain sad for longer periods of time. This is encouraging towards further exploration of dynamic models to understand such complex behaviors.

7. Conclusions and Future Work

Human behavior is highly dynamic. By exploiting the context of human behavior, we show that our proposed dynamic models are more robust in behavioral classification than static models. Further, as an intermediate output our model provides a trove of information about local behavioral patterns that can be exploited in the future for finer analysis. This can be useful to domain experts, for example, to detect triggers of certain behavioral expressions such as reactivity, thus supporting their diagnosis and treatment.

In the future, we would like to model the interpersonal dynamics in addition to temporal dynamics. We intend to employ a Dynamic Bayesian Network to infer the state transition process between couples based on both expressed and perceived stimuli. Moreover, in our parallel work [25], we have reduced the original 33 behavioral codes to a lower dimensional space. We would like to exploit the cross-behavior information and the reduced dimensionality to design better behavioral classification methods.

8. References

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

[2] M. Pantic, A. Nijholt, A. Pentland, and T. S. Huang, "Human-centred intelligent human computer interaction (hci^2): how far are we from attaining it?" *International Journal of Autonomous and Adaptive Communications Systems*, vol. 1, no. 2, pp. 168–187, 2008.

[3] B. Schuller, M. Wimmer, L. Mosenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?" in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4501–4504.

[4] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction*, 2011, pp. 415–424.

[5] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, 2012, pp. 449–456.

[6] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, May 2013.

[7] P. G. Georgiou, M. P. Black, A. Lammert, B. Baucom, and S. S. Narayanan, ""That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Proceedings of Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, Oct. 2011.

[8] P. G. Georgiou, M. P. Black, and S. S. Narayanan, "Behavioral signal processing for understanding (distressed) dyadic interactions: Some recent developments," in *Third International Workshop on Social Signal Processing, ACM Multimedia*, Scottsdale, AZ, 2011, pp. 7–12.

[9] M. Black, A. Katsamanis, C.-C. Lee, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Automatic classification of married couples' behavior using audio features." in *Proceedings of Interspeech*, 2010, pp. 2030–2033.

[10] A. Metallinou, R. B. Grossman, and S. Narayanan, "Quantifying atypicality in affective facial expressions of children with autism spectrum disorders," in *IEEE International Conference on Multimedia and Expo*, 2013, pp. 1–6.

[11] A. Christensen, D. C. Atkins, S. Berns, J. Wheeler, D. H. Baucom, and L. E. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples." *Journal of Consulting and Clinical Psychology*, vol. 72, no. 2, pp. 176–191, 2004.

[12] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features," *Speech Communication*, vol. 55, no. 1, pp. 1–21, Jan. 2013.

[13] A. Katsamanis, J. Gibson, M. P. Black, and S. S. Narayanan, "Multiple instance learning for classification of human behavior observations," in *Affective Computing and Intelligent Interaction*, 2011, pp. 145–154.

[14] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *The Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[15] S. N. Chakravarthula, R. Gupta, B. Baucom, and P. Georgiou, "A language-based generative model framework for behavioral analysis of couples' therapy," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2015.

[16] M. Collins and N. Duffy, "New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 263–270.

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

[18] J. Jones and A. Christensen, "Couples interaction study: Social support interaction rating system," University of California, Los Angeles, Technical manual, 1998.

[19] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, Mar. 2011.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, 2010, pp. 1459–1462.

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

[22] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[23] K. Murphy, "Hidden markov model toolbox for matlab," *online at http://www.cs.ubc.ca/˜murphyk/Software/HMM/hmm.html*, 2005.

[24] P. Borkenau, N. Mauer, R. Riemann, F. M. Spinath, and A. Angleitner, "Thin slices of behavior as cues of personality and intelligence." *Journal of Personality and Social Psychology*, vol. 86, no. 4, pp. 599–614, 2004.

[25] M. Nasir, B. Baucom, P. Georgiou, and S. S. Narayanan, "Redundancy analysis of behavioral coding for couples therapy and improved estimation of behavior from noisy annotations," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2015.