Automatic Classification of Married Couples' Behavior using Audio Features

Matthew Black[1], *Athanasios Katsamanis*[1], *Chi-Chun Lee*[1], *Adam C. Lammert*[1],
Brian R. Baucom[2], *Andrew Christensen*[3], *Panayiotis G. Georgiou*[1], *Shrikanth Narayanan*[1,2]

[1]Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA
[2]Department of Psychology, University of Southern California, Los Angeles, CA, USA
[3]Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA
http://sail.usc.edu[1], baucom@usc.edu[2], christensen@psych.ucla.edu[3]

Abstract

In this work, we analyzed a 96-hour corpus of married couples spontaneously interacting about a problem in their relationship. Each spouse was manually coded with relevant session-level perceptual observations (e.g., level of blame toward other spouse, global positive affect), and our goal was to classify the spouses' behavior using features derived from the audio signal. Based on automatic segmentation, we extracted prosodic/spectral features to capture global acoustic properties for each spouse. We then trained gender-specific classifiers to predict the behavior of each spouse for six codes. We compare performance for the various factors (across codes, gender, classifier type, and feature type) and discuss future work for this novel and challenging corpus.

Index Terms: behavioral signal processing, human behavior analysis, couples therapy, prosody, emotion recognition

1. Introduction

Several fields in psychology depend critically on perceptual judgments made by people. For example, diagnoses of social disorders (e.g., autism) and many types of social therapies (e.g., couples therapy), require 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. In family studies research and practice, psychologists rely on a variety of established coding standards [1].

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 [1] [2].

Technology has the potential to greatly help with coding audio-visual data. Certain measurements are difficult or impossible for humans to do, such as accurately tracking the pitch of a speaker or quantifying a person's movement. Computers are much better-suited to extract these so-called low-level descriptors (LLDs) of human behavior [3]. Human behavioral signal processing involves using signal processing methods and machine learning algorithms to extract human-centered information from audio-video signals, including social cues [2], affect and emotions [4] [5] [6], and intent [7]. Rather than relying on multiple humans to laboriously code audio-video data, the idea here is to use LLDs and mid-level signal representations to estimate perceptual judgments at possibly multiple granularities. The advantage of using computers is that it could provide a consistent way to automatically quantify aspects of human behavior. The technology could also potentially be adapted from one domain to another.

In this paper, we analyze recordings of a husband and wife discussing a problem they are having with their relationship. The spontaneous sessions were manually labeled with a number of session-level perceptual codes (e.g., global negative affect for each spouse). Affective states and intentions are often portrayed vocally [8], and verbal cues have been found to be relevant in the context of marital conflicts [9]. Our goal in this paper is to learn the perceptual codes directly from the audio signal to demonstrate the predictive power of objective signal-based cues. Section 2 describes the corpus. Section 3 discusses the acoustic features we extracted, and Section 4 displays the classification results. Section 5 provides a discussion, and we conclude in Section 6 with future work.

2. Corpus

The corpus we are using consists of audio-video recordings of couples (wife and husband) during real problem-solving dyadic interactions. The data was collected as part of a longitudinal study at the University of California, Los Angeles and the University of Washington. 134 seriously and chronically distressed married couples received couples therapy for one year. Participants in the study ranged from 22 to 72 years old, with a median age for men of 43 years (SD = 8.8) and a median age for women of 42 years (SD = 8.7). They were, on average, college-educated (median level of education for both men and women was 17 years, SD = 3.2). The sample was largely Caucasian (77%), with 8% African American, 5% Asian or Pacific Islander, 5% Latino/Latina, 1% Native American, and 4% Other. Couples were married an average of 10.0 years (SD = 7.7) [10].

As part of the study, the couples participated in sessions where they discussed a problem in their relationship with no therapist or research staff present. The couple talked for ten minutes about the wife's chosen topic and ten minutes about the husband's chosen topic; these sessions were considered separate. Each couple's problem-solving interactions were recorded at three points in time: before the therapy sessions began, 26 weeks into therapy, and two years after the therapy sessions finished. In total, we have 96 hours of data across 574 sessions.

The audio-video data consist of a split-screen video (704x480 pixels, 30 fps) and a single channel of far-field audio. Since the data was originally only intended for manual coding, the recording conditions were not ideal for automatic analysis; the video angles, microphone placement, and background noise varied across couples and across time periods. We also have access to word transcriptions, in which the speaker was labeled as

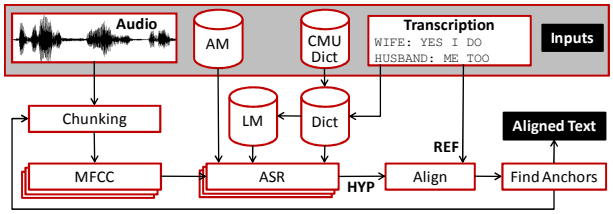

Figure 1: *Block diagram of the automatic alignment procedure [13]. Generic acoustic models (AM) and session-specific language models (LM) are used to run ASR on the audio file. Anchor regions are accepted if aligned portions between the reference transcript (REF) and ASR transcript (HYP) meet 4 heuristics. The process is then iterated between anchor regions.*

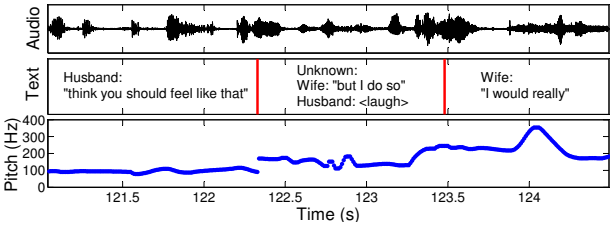

Figure 2: *Example of automatic alignment and processed pitch signal. In this particular segment the middle portion (labeled "Unknown") was not aligned due to overlapped speech (the husband was laughing while the wife was speaking).*

well (husband/wife). The transcripts lack detailed annotations such as timing and speech overlap indications.

Both spouses were evaluated with 33 session-level codes from two coding schemes. The Social Support Interaction Rating System (SSIRS) consists of 20 codes that measure the emotional component of the interaction and the topic of conversation [11]. The 13 codes in the Couples Interaction Rating System 2 (CIRS2) were specifically designed for conversations involving a problem in a relationship [12]. Both coding manuals were designed to have evaluators watch the entire session and provide session-level ratings of *each* spouse's overall behavior on an integer scale from 1 to 9; utterance- and turn-level ratings were not obtained. Three to four student evaluators coded each session, producing one set of 33 codes for each spouse. All evaluators underwent a training period to give them a sense for what was "typical" behavior and to help standardize the coding process.

Due to low inter-evaluator agreement for some codes and high correlation between some of the codes, we chose to analyze six codes for this paper. Two of the codes were from CIRS2: level of acceptance toward the other spouse (abbreviated "acc") and level of blame ("bla"), and the other four codes were from SSIRS: global positive affect ("pos"), global negative affect ("neg"), level of sadness ("sad"), and use of humor ("hum"). 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 *and* negative affect if they display both often enough. Table 1 shows the correlation between the six codes and between the wife's and husband's scores, and the inter-evaluator agreement for each code.

Using the transcripts, we created word and speaker-turn alignments using a recursive automatic speech recognition (ASR) technique based on [13]. Figure 1 shows a block diagram. After the algorithm converged, each session was split into

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

Table 1: *Pearson's correlation between codes/spouse and inter-evaluator agreement for the six codes we analyzed.*

speaker-homogeneous regions (wife speech, husband speech) and unknown regions in which alignment was not achieved (Figure 2). We were able to segment more than 60% of the sessions' words into speaker-homogeneous regions for 293 of the 574 sessions (which included 83 unique husband/wife pairs). The other sessions were deemed too noisy to achieve good segmentation using this automatic alignment technique and are ignored for the remainder of this paper.

3. Feature Extraction

Prosodic cues (e.g., pitch, energy, timing) have been shown to be relevant in psychology literature [8] [9], and prosodic and spectrum-based audio features have been used extensively for various human-centered learning tasks [3] [4] [5] [6] [7]. We explore the use of several acoustic features in predicting high-level perceptual judgments about the couples' behavior.

We computed the speaking rate for each aligned word from the automatic segmentation output (in units of words/s and letters/s). We next separated speech from non-speech regions by running a voice-activity detector (VAD), trained on a 30-second sample from one held-out session [14]. We extracted two first-order Markov chain features from the VAD output: the transition probabilities to and from speech and non-speech states. We also extracted the lengths of each speech and non-speech segment to use as a LLD for later feature extraction.

In addition to this VAD stream, we extracted the following LLDs across each *speech* region every 10 ms using a 25 ms Hamming window: pitch, root-mean-square energy, harmonics-to-noise ratio (HNR), voice quality (computed as the zero-crossing rate of the autocorrelation function), 13 MFCCs, 26 magnitude of Mel-frequency band (MFB) features, and magnitude of the spectral centroid and spectral flux. All LLDs except pitch were extracted with openSMILE [15].

Pitch estimates were made with Praat [16] using an autocorrelation method, with minimum and maximum pitch values of 65 Hz and 500 Hz, respectively. The resulting pitch signal was then passed through an algorithm that attempted to fix instances of pitch halving/doubling across unvoiced regions by detecting large jumps in the pitch difference vector and was subsequently median filtered and linearly interpolated (with no interpolation across speaker-change points and regions detected as non-speech by the VAD). See Figure 2 for an example.

In addition to this processed pitch signal, we also computed two normalized pitch streams by: 1) subtracting the mean pitch of the speaker for that frame (wife, husband, or unknown), and 2) performing a similar normalization on a logarithmic scale (Eq. 1). The mean pitch value was computed across the whole session using the automatic segmentation results; unknown regions were treated as coming from one unknown "speaker."

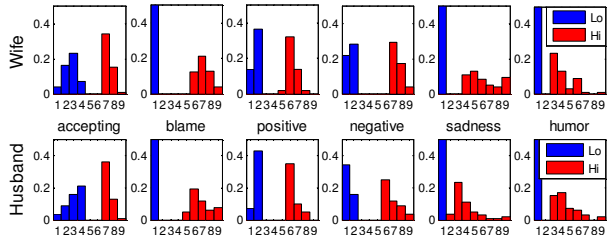

Figure 3: *Normalized histograms of code scores for the wife (top) and husband (bottom). Scores in blue are in the bottom 50 (20%), and scores in red are in the top 50 (20%).*

$$F_{0_{\text{norm1}}} = F_0 - \mu_{F_0}, \qquad F_{0_{\text{norm2}}} = \frac{\log\left(\frac{F_0}{\mu_{F_0}}\right)}{\log(2)} \tag{1}$$

Each session was split into three "domains": wife (speaker-homogeneous regions where the wife was the speaker), husband (speaker-homogeneous regions where the husband was the speaker), and independent (alignment-independent regions that included the whole session). We produced the final set of features by extracting 13 functionals across each domain for each LLD: mean, standard deviation, skewness, kurtosis, range, minimum, minimum location (normalized by the session length), maximum, maximum location (normalized), first quartile, median, third quartile, and interquartile range. These 2007 features capture global acoustic properties of the spouses/interaction in each session. Note that delta and delta-delta features for the LLDs were not included in the final feature set, since they offered no improvement in initial experiments.

4. Classification Results

We framed this as a binary classification task (as in [7]), rather than predicting the codes in a continuous manner. We analyzed only the sessions that had mean scores (averaging across evaluators) that fell in the top 50 and bottom 50 of the score range, corresponding to approximately the top 20% and bottom 20% (Figure 3). Thus, our goal was to separate the two extremes automatically. We trained wife and husband models separately (12 binary classifiers total), and the percentage of misclassified sessions was the chosen error metric, with a voting baseline performance of 50% error.

We compared two classifiers: 1) support vector machines (SVM) with linear kernel, and 2) Fisher's linear discriminant analysis (LDA), where we chose the discriminant hyperplane to lie exactly between the projection of the class means. We used leave-one-*couple*-out cross-validation to separate train and test sets; we did not use leave-one-*session*-out cross-validation since some couples had more than one session in the top and/or bottom 50. The SVM was trained using all features; Fisher's LDA used forward feature selection, where the best feature was iteratively selected if it improved average classification accuracy over ten folds of the train set. To reduce the chance of poor feature selection, all features that were highly uncorrelated (magnitude of Pearson correlation coefficient less than 0.3) with the class labels at each cross-validation were disregarded.

Table 2 shows the classification error for each code/gender/classifier combination. All errors equal to or less than 31% are significantly lower than the voting baseline performance of 50% with $p < 0.05$, using McNemar's Test.

Classifier	Spouse	*acc*	*bla*	*pos*	*neg*	*sad*	*hum*	AVG
linear SVM	Wife	31	**33**	23	**26**	**34**	**47**	32.3
(all 2007 features)	Husband	27	**32**	33	**24**	53	25	32.4
Fisher's LDA	Wife	**20**	35	**20**	28	**34**	**47**	**30.7**
(forward selection)	Husband	**21**	39	**31**	27	**35**	**21**	**29.0**

Table 2: *Number (percentage) of sessions misclassified for each code and spouse when using various classifiers.*

5. Discussion

Fisher's LDA performed better than the SVM classifier for four of the codes and had a lower average classification error; the LDA method most likely benefitted from the forward feature selection. We got the best performance when classifying the code, "acceptance of other spouse," and the worst results when trying to predict the wife's "use of humor." We were surprised by the large disparity between the performance for "acceptance" and "blame," since they were highly anti-correlated codes; however, the coding manual for blame concentrated on lexical cues, so the coders may have been more affected by *what* the spouses said, not *how* they said it. As a result, just the acoustic features may not be good at discriminating high blame from low blame.

An average of 3.4 features (*SD*=0.81) were selected at each fold by the LDA classifier. When training the wife models, the domain break-down on selected features was: Independent=56%, Wife=32%, Husband=12%. For the husband models: Independent=41%, Husband=37%, Wife=22%. The LLD break-down on selected features was: Pitch=30%, MFCC=26%, MFB=26%, VAD=12%, Rate=4%, Other=2%. Further analyzing the selected pitch features, 47% were $F_{0_{\text{norm1}}}$ features (Eq. 1), 36% were $F_{0_{\text{norm2}}}$ features (Eq. 1), and 17% were from the unnormalized pitch signal.

To gain more insight into the predictive power of each domain/LLD, we ran a second set of experiments by re-training Fisher's LDA on features from a single domain/LLD. Tables 3 and 4 show these results. Looking at individual domains, we see that performance decreased for the husband when only wife domain features are used (and vice versa). This makes sense, since each spouse was rated separately, and coders probably based their ratings more on the regions in which the spouse being rated was talking. However, it is interesting to note that the performance does not severely drop in these mismatched conditions. This is in part due to the fact that there is positive correlation between the spouses' scores (Table 1). Another interesting trend is the good performance when using alignment-independent features only, even outperforming the wife predictions when using only wife domain features. This is probably due to the fact that the session is inherently interactive, and the alignment-independent features capture that better than the husband-only and wife-only domain features. Thus, context is important and should be better modeled in the future.

Looking at individual LLDs, we see that no single LLD has an average code performance better than the case when all features were used, which is to be expected. However, for some codes/genders, performance is better when trained on a single LLD. This means that there is still plenty of room to improve with our learning methods through better feature selection and dimensionality reduction. The relative goodness of the pitch features also implies that more emphasis needs to be placed on feature normalization techniques, so that the features are more generalizable from session-to-session and couple-to-couple.

Domain	N	Spouse	*acc*	*bla*	*pos*	*neg*	*sad*	*hum*	AVG
Wife	669	Wife	23	**31**	26	24	**34**	58	32.7
		Husband	38	44	27	33	56	46	40.7
Husband	669	Wife	36	37	38	46	59	**39**	42.5
		Husband	**25**	**35**	28	29	**33**	**31**	**30.2**
Independent	669	Wife	**22**	44	**18**	**23**	35	43	**30.8**
		Husband	35	39	**25**	**28**	35	51	35.5

Table 3: *Number (percentage) of sessions misclassified when using features from a single domain (using Fisher's LDA classifier with forward feature selection).*

LLD	N	Spouse	*acc*	*bla*	*pos*	*neg*	*sad*	*hum*	AVG
Pitch	117	Wife	36	37	38	46	59	39	42.5
		Husband	25	35	**28**	29	**33**	**31**	**30.2**
VAD	96	Wife	35	42	40	39	**32**	66	42.3
		Husband	42	40	42	27	**33**	50	39.0
Energy	39	Wife	32	40	36	34	64	49	42.5
		Husband	34	**25**	37	28	51	33	34.7
Rate	78	Wife	37	**35**	56	47	44	**38**	42.8
		Husband	30	53	31	32	43	43	38.7
HNR	39	Wife	**29**	44	37	37	38	46	38.5
		Husband	36	30	32	29	57	61	40.8
Voice Quality	39	Wife	44	**35**	52	40	47	44	43.7
		Husband	41	**25**	46	30	56	39	39.5
MFCC	507	Wife	40	36	27	37	43	48	38.5
		Husband	29	33	38	29	48	43	36.7
MFB	1014	Wife	**29**	53	**23**	**33**	38	47	**37.2**
		Husband	**19**	46	30	30	56	41	37.0
Spectral Centroid	39	Wife	42	39	46	64	58	56	50.8
		Husband	46	46	50	33	62	43	46.7
Spectral Flux	39	Wife	34	43	28	46	41	47	39.8
		Husband	41	40	44	**26**	50	32	38.8

Table 4: *Number (percentage) of sessions misclassified when using features from a single LLD (using Fisher's LDA classifier with forward feature selection).*

6. Conclusion & Future Work

This work represents an initial analysis of a novel corpus consisting of real couples interacting about problems in their relationship. We showed that we could train binary classifiers using only audio features that separated spouses' behavior significantly better than chance for four of the six codes we examined. This is a challenging learning problem due to the absence of utterance/turn-level behavioral codes and the inherent complexity of the dyadic interaction. This type of research is important, since there is a dearth of work in the psychology literature that focuses on objective signal-based cues for human behavioral analysis.

In the continuation of this work, we will investigate fusion of the audio features with lexical features, which are also showing promising results. Future work will incorporate dynamic modeling that captures within-utterance variations and cross-turn transitions and explores the use of saliency detection. Additionally, we would also like to incorporate expert information into this learning framework, which could help in a number of ways, such as informing a lower-dimensional and focused feature set. Lastly, we are currently collecting a new database of dyadic interactions using a 10-HD camera, 15-microphone acquisition with the aid of a motion capture system. This new corpus will allow for multimodal analysis of dyadic interactions [17].

7. Acknowledgments

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

8. References

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

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

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

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

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

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

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

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

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

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

[11] J. Jones and A. Christensen, *Couples interaction study: Social support interaction rating system*, University of California, Los Angeles, 1998. [Online]. Available: http://christensenresearch.psych.ucla.edu/

[12] C. Heavey, D. Gill, and A. Christensen, *Couples interaction rating system 2 (CIRS2)*, University of California, Los Angeles, 2002. [Online]. Available: http://christensenresearch.psych.ucla.edu/

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

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

[15] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR–Introducing the Munich open-source emotion and affect recognition toolkit," *Proc. IEEE ACII*, 2009.

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

[17] V. Rozgić, B. Xiao, A. Katsamanis, B. Baucom, P. G. Georgiou, and S. Narayanan, "A new multichannel multimodal dyadic interaction database," in *Proc. Interspeech*, 2010.