

Details of the Classification System

Movement artifacts were automatically detected in EDA [1] and were visually inspected and manually revised. After cleaning, automatic detection of skin conductance responses (SCR) was further performed [2]. SCRs and skin conductance levels were averaged over each hourly period to yield 2-dimensional EDA features. QRS complexes were further automatically detected in ECG [3]. EDA synchrony was computed as an index of signal similarity based on a joint sparse representation approach with appropriately designed EDA-specific dictionaries [4]. The analysis window of the sparse representation was 15 min to account for adequate signal variability. These 15 min synchrony indices were averaged over each hour resulting in the final 1-dimensional synchrony score.

Audio processing involved voice activity detection (VAD) to automatically chunk continuous audio streams into segments of speech or non-speech. Our VAD system exploits the short- and long-term spectral information of the audio signal with a multilayer perceptron classifier [5]. Speaker clustering and gender identification were used to automatically assign a gender to each speech segment by taking into account acoustic, prosodic and voice quality information with a Gaussian mixture model framework [6], [7].

Classification of the one-dimensional, theoretically-driven features (Task 1) was performed using a binary decision tree because of its efficiency and ability to capture non-linearity between the outcome and the feature space. In order to overcome the high dimensional and highly correlated feature spaces during the unimodal (Task 2) and multimodal (Task 3) classification, we employed an autoassociative neural network, also called “autoencoder.” Autoencoders have been used in a variety of tasks for unsupervised dimensionality reduction and are able to capture the non-linear associations of the input space and remove the underlying redundancies [8, 9]. The autoencoder aims to learn a low-dimensional representation of the input data by performing identity mapping between the input and output layers, thus minimizing the error between original and reconstructed data. Given the available number of samples in our task, the autoencoder consisted of one input, one output and three hidden layers all fully connected (Figure 1 in the manuscript). The middle layer contains the final features of interest, also called “bottleneck” features. The dimensionality of the bottleneck features was empirically fixed to be half the dimensionality of the original input space. The component values yielding from the autoencoder were fed into the binary decision tree classifier. The cost C of misclassifying a non-conflict sample as conflict through the decision tree was tuned with a nested cross validation [10], in which the inner loop conducts a grid search to find the optimal parameter among $C = 0.05, 0.10, \dots, 0.95, 1$, where the default value of equal misclassification cost between the two classes is one. We used a leave-one-couple-out cross-validation setup for all classification experiments.

Because of the disproportionate number of samples between the two classes of interest, evaluation of the classification task was performed through unweighted accuracy (UA) [11]. This was computed as the average percentage of the number of accurately recalled hourly instances for each conflict/non-conflict class. Significance of the resulting UAs compared to by-chance accuracy (50%) was statistically determined through a t -test. Comparison of the multimodal indices to the couples’ self-reported MQI was also performed through a t -test for the corresponding UAs (Task 4). We computed sensitivity and specificity to provide separate measures of accuracy for the conflict and non-conflict samples. In order to check the performance of the proposed system in various discrimination thresholds, we further report the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The ROC operating points were determined by varying the misclassification cost C between 0 and 1 (with 0.05 step).

[1] T. Chaspari, A. Tsiartas, L. I. Stein, S. A. Cermak, and S. S. Narayanan, “Sparse representation of electrodermal activity with knowledge-driven dictionaries,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 960–971, 2015.

[2] M. Benedek and C. Kaernbach, “Decomposition of skin conductance data by means of nonnegative deconvolution,” *Psychophysiology*, vol. 47, no. 4, pp. 647–658, 2010.

[3] “The BioSig Project,” <http://biosig.sourceforge.net/>.

[4] T. Chaspari, B. Baucom, A. C. Timmons, A. Tsiartas, L. Borofsky Del Piero, K. J. W. Baucom, P. Georgiou, G. Margolin, and S. S. Narayanan, “Quantifying EDA synchrony through joint sparse representation: A case-study of couples’ interactions,” In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2015.

[5] M. Van Segbroeck, A. Tsiartas, and S. Narayanan, “A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice.,” in *Proc. INTER-SPEECH*, 2013, pp. 704–708.

- [6] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.
- [7] N. Kumar, M. Nasir, P. Georgiou, and S. Narayanan, "Robust multi-channel gender classification from speech in movie audio.," in *Proc. INTERSPEECH*, 2016.
- [8] M. Scholz, "Validation of nonlinear PCA," *Neural Processing Letters*, vol. 36, no.1, pp. 21-30, 2012.
- [9] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," In *Proc. International Conference on Artificial Neural Networks*, 2011
- [10] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.
- [11] H. L. Wagner, "On measuring performance in category judgment studies of nonverbal behavior," *Journal of Nonverbal Behavior*, vol. 17, no. 1, pp. 3–28, 1993.

TABLE 1A
 CONFLICT CLASSIFICATION MEASURES FROM SELF-REPORTED MQI, EDA, ECG, AND EDA SYNCHRONY FEATURES (TASK 1)

Feature	Female Partner				Male Partner			
	UA	Sensitivity	Specificity	AUC	UA	Sensitivity	Specificity	AUC
Stressed	59.8*	32.4	87.1	0.47	60.5**	48.4	72.6	0.58
Happy	66.8**	43.4	90.3	0.66	60.6*	29.2	92.0	0.50
Sad	65.5**	38.1	92.8	0.55	59.9*	27.7	92.0	0.50
Nervous	59.0*	30.4	87.7	0.43	55.0	24.5	85.4	0.43
Angry	69.5**	54.3	84.6	0.67	69.2**	46.8	91.7	0.54
Close	63.5**	77.2	49.8	0.59	60.6**	24.4	96.8	0.51
SCL	49.3	17.8	80.9	0.48	54.5	54.4	54.7	0.49
SCR	52.8	27.7	77.9	0.48	53.9	57.6	50.1	0.55
IBI	61.0**	51.8	70.1	0.56	61.4**	67.3	55.6	0.60
HRV	51.6	48.7	54.6	0.51	62.8**	53.0	72.6	0.62
EDA Synchrony	55.3	31.8	78.7	0.45	56.6	29.2	84.0	0.48

Note. Percentages of UA, sensitivity, specificity, and AUC for individual features. UA = unweighted accuracy, AUC = area under the curve, SCL = skin conductance level, SCR = skin conductance response, IBI = interbeat interval, HRV = heart rate variability, MQI = mood and quality of interactions, EDA = electrodermal activity, ECG = electrocardiogram, * = $p < 0.05$, ** = $p < 0.01$ (UA significantly higher than 50% chance).

TABLE 1B
 CONFLICT CLASSIFICATION MEASURES FROM CONTEXT AND INTERACTION FEATURES (TASK 1)

Feature	Female Partner				Male Partner			
	UA	Sensitivity	Specificity	AUC	UA	Sensitivity	Specificity	AUC
Consume Caffeine	52.6	5.3	100	0.36	52.6	5.3	100	0.37
Consume Alcohol	52.6	5.3	100	0.38	52.6	5.3	100	0.34
Consume Tobacco	52.6	5.3	100	0.41	52.6	5.3	100	0.37
Consume Other Drugs	52	5.3	98.8	0.41	52.6	5.3	100	0.42
Physical Activity	52.6	5.3	100	0.32	52.6	6.3	98.9	0.38
Body Temperature	47.1	18.2	76.1	0.41	53	68.2	37.7	0.51
Activity Count	49.5	44.2	54.9	0.48	53.5	32.9	74.2	0.44
Driving	60.1*	39.6	80.7	0.56	59.8*	37	82.6	0.55
GPS Distance	56.3	69.1	43.6	0.52	59.6*	55.1	64.1	0.5
Together	52.6	5.3	100	0.4	52.6	5.3	100	0.39
Interacting	52.6	5.3	100	0.36	52.6	5.3	100	0.43
Phone Interview	52.6	5.3	100	0.4	52.6	5.3	100	0.39
Interacting with Others	52.6	5.3	100	0.44	55	54.9	55.1	0.49

Note. Percentages of UA, sensitivity, specificity, and AUC for individual features. UA = unweighted accuracy, AUC = area under the curve, GPS = global positioning system, * = $p < 0.05$, ** = $p < 0.01$ (UA significantly higher than 50% chance).

TABLE 1C
CONFLICT CLASSIFICATION MEASURES FROM LINGUISTIC FEATURES (TASK 1)

Feature	Female Partner				Male Partner			
	UA	Sensitivity	Specificity	AUC	UA	Sensitivity	Specificity	AUC
Number of words	48.1	52.8	43.5	0.43	58.6**	65.8	51.5	0.56
Words with more than 6 letters	55.3	35	75.6	0.52	52.1	60.0	44.3	0.53
Words in LIWC dictionary	51.9	29.9	73.9	0.49	62.8**	76.2	49.4	0.59
Function words	58.6*	56.8	60.5	0.53	59.3*	33.2	85.5	0.55
Pronouns	56.4	62.5	50.4	0.51	51.4	18.9	83.8	0.48
Personal pronouns	50.1	16.8	83.4	0.48	52.8	21.3	84.2	0.45
First-person singular pronouns	49.0	60.6	37.3	0.44	53.5	52.3	54.7	0.49
Second-person plural pronouns	48.4	60.4	36.3	0.43	56.6*	24.8	88.4	0.53
Second-person pronouns	51.2	54.4	47.9	0.5	48.6	54.5	42.8	0.41
Third-person singular pronouns	52.4	72.0	32.9	0.43	52.1	72.7	31.5	0.47
Third person plural pronouns	57.7	26.4	89.0	0.51	52.6	14.3	90.8	0.47
Impersonal pronouns	57.4	59.6	55.1	0.55	58.0	68.5	47.4	0.49
Articles	52.4	56.9	47.8	0.52	58.3*	55.5	61.0	0.51
Verbs	57.1*	64.3	49.8	0.56	60.9*	60.2	61.6	0.59
Auxiliary verbs	55.9	53.1	58.7	0.55	56.9	65.7	48.2	0.54
Past tense	53.2	65.9	40.5	0.48	48.5	16.9	80.1	0.42
Present tense	57.5	68.6	46.4	0.57	52.6	47.2	58.0	0.48
Future tense	55.7	27.2	84.2	0.49	60.3**	79.2	41.3	0.54
Adverbs	48.5	13.0	84.0	0.43	48.7	13.3	84.1	0.39

Prepositions	52.7	55.8	49.5	0.48	50.8	64.2	37.3	0.42
Conjunctions	53.8	27.5	80.1	0.49	57.1	54.2	60.0	0.55
Negations	51.1	16.6	85.6	0.44	56.4	74.3	38.4	0.50
Quantifiers	54.1	63.5	44.6	0.51	51.1	65.3	36.9	0.49
Numbers	50.5	16.0	85.1	0.41	56.9	74.5	39.4	0.53
Swear words	55.3	19.8	90.8	0.41	55.8	29.6	82.1	0.48

Note. Percentages of UA, sensitivity, specificity, and AUC for individual features. UA = unweighted accuracy, AUC = area under the curve, * = $p < 0.05$, ** = $p < 0.01$ (UA significantly higher than 50% chance). With the exception of the number of words, values per category were calculated as a proportion of total words spoken. See LIWC manual for definitions of specific word categories [1].

[1] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth, "The development and psychometric properties of LIWC 2007," *LIWC.net*, 2007.

TABLE 1D
 CONFLICT CLASSIFICATION MEASURES FROM PSYCHOLOGICAL FEATURES (TASK 1)

Feature	Female Partner				Male Partner			
	UA	Sensitivity	Specificity	AUC	UA	Sensitivity	Specificity	AUC
Social processes	50.1	59.2	40.9	0.47	54.8	57.3	52.3	0.48
Family words	52.5	25.8	79.2	0.51	54.4	14.7	94.0	0.47
Friend words	55.9	18.3	93.5	0.52	52.6	10.5	94.6	0.48
Humans words	54.3	45.5	63.1	0.5	50.9	16.1	85.8	0.46
Affective processes	49.9	22.3	77.5	0.43	52.3	60.9	43.6	0.44
Positive emotion	54.8	68.9	40.6	0.47	55.3	55.9	54.7	0.47
Negative emotion	62.3**	77.4	47.2	0.53	55.9	60.1	51.6	0.51
Anxiety	54.1	18.9	89.3	0.48	53.2	13.7	92.7	0.44
Anger words	55.5	55.0	55.9	0.48	55.7	71.3	40.0	0.50
Sad	54.3	23.0	85.5	0.49	57.7*	21.7	93.8	0.49
Cognitive processes	57.7	55.8	59.6	0.59	50.2	24.4	75.9	0.45
Insight	53.3	28.2	78.4	0.44	59.1**	55.6	62.6	0.53
Causation	53.1	22.1	84.2	0.45	48.8	16.0	81.7	0.38
Discrepancy	56.6	61.7	51.6	0.50	58.1*	55.0	61.2	0.52
Tentative	50.2	26.1	74.3	0.45	57.5**	33.1	82.0	0.52
Certainty	52.4	23.8	81.1	0.43	58.4*	57.7	59.1	0.54
Inhibition	56.7	32.5	80.9	0.50	61.4**	56.3	66.5	0.57
Inclusive	50.7	13.4	88.0	0.45	53.3	66.8	39.8	0.48
Exclusive	53.8	63.5	44.1	0.49	51.1	67.3	34.9	0.46

Perceptual processes	59.1*	67.1	51.1	0.54	56.4	25.4	87.4	0.53
See	57.6	32.5	82.7	0.52	57.0*	73.2	40.8	0.50
Hear	50.5	11.3	89.6	0.44	51.6	54.0	49.1	0.47
Feel	55.4	24.8	86.0	0.43	59.7**	70.1	49.4	0.55
Biological processes	50.3	13.7	86.9	0.45	52.1	65.6	38.6	0.47
Body	54.3	17.6	91.0	0.50	52.1	71.4	32.9	0.42
Health	52.7	20.6	84.9	0.40	50.3	9.4	91.2	0.39
Sexual	50.9	80.3	21.4	0.49	53.0	18.2	87.7	0.44
Ingestion	50.8	16.9	84.8	0.42	57.1	24.9	89.3	0.45
Relativity	45.5	10.1	80.9	0.36	56.0	49.9	62.2	0.50
Motion	55.7*	25.1	86.3	0.50	63.7**	78.3	49.2	0.63
Space	55.8	23.1	88.5	0.52	54.6	63.9	45.3	0.47
Time	54.3	64.3	44.3	0.49	56.1	51.5	60.7	0.56

Note. Percentages of UA, sensitivity, specificity, and AUC for individual features. UA = unweighted accuracy, AUC = area under the curve, * = $p < 0.05$, ** = $p < 0.01$ (UA significantly higher than 50% chance). Values per category were calculated as a proportion of total words spoken. See LIWC manual for definitions of specific word categories [1].

[1] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth, "The development and psychometric properties of LIWC 2007," *LIWC.net*, 2007.

TABLE 1E
CONFLICT CLASSIFICATION MEASURES FROM PERSONAL AND PARALINGUISTIC FEATURES (TASK 1)

Feature	Female Partner				Male Partner			
	UA	Sensitivity	Specificity	AUC	UA	Sensitivity	Specificity	AUC
Work	52.1	21.1	83.2	0.42	53.6	72.4	34.8	0.49
Achievement	56.1	24.3	87.9	0.47	57.3	28.3	86.4	0.50
Leisure	58.1	30.8	85.3	0.49	57.9*	29	86.8	0.53
Home	55.1	19.2	90.9	0.45	52.0	16.3	87.7	0.39
Money	52.0	11.9	92.2	0.47	53.4	77	29.9	0.51
Religion	51.7	7.2	96.1	0.36	50.9	8.3	93.5	0.40
Death	57.5	15.1	100	0.42	51.6	5.3	98.0	0.39
Assent	55.7	61.9	49.4	0.51	56.5*	63.6	49.4	0.54
Non-fluencies	53.4	68.7	38.0	0.48	55.8	78.3	33.3	0.50
Filler	58.6*	60.7	56.4	0.56	54.7	50.8	58.6	0.48

Note. Percentages of UA, sensitivity, specificity, and AUC for individual features. UA = unweighted accuracy, AUC = area under the curve, * = $p < 0.05$, ** = $p < 0.01$ (UA significantly higher than 50% chance). Values per category were calculated as a proportion of total words spoken. See LIWC manual for definitions of specific word categories [1].

[1] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth, "The development and psychometric properties of LIWC 2007," *LIWC.net*, 2007.

TABLE 1F
 CONFLICT CLASSIFICATION MEASURES FROM AUDIO FEATURES (TASK 1)

Feature	Female Partner				Male Partner			
	UA	Sensitivity	Specificity	AUC	UA	Sensitivity	Specificity	AUC
Mean intensity	49.7	55.5	43.9	0.43	57.4*	28.3	86.4	0.54
Intensity slope	59.3*	70.4	48.2	0.58	48.3	52.7	43.9	0.42
Maximum intensity	58.8*	55.8	61.8	0.53	55.7	51.8	59.6	0.50
Intensity standard deviation	59.6**	71.1	48.1	0.53	54.4	44.3	62.4	0.49
Mean F0	53.7	10.5	96.9	0.46	57.2*	45.1	69.3	0.50
F0 slope	59.5*	34.2	84.8	0.48	52.5	10.9	94.1	0.36
Maximum F0	55.2	23.1	87.2	0.40	53.7	8.2	99.1	0.41
F0 standard deviation	52.6	10.5	94.7	0.42	52.6	9.2	9.6	0.35

Note. Percentages of UA, sensitivity, specificity, and AUC for individual features. UA = unweighted accuracy, AUC = area under the curve, F0 = fundamental frequency, * = $p < 0.05$, ** = $p < 0.01$ (UA significantly higher than 50% chance).

TABLE 2
CONFLICT CLASSIFICATION MEASURES FROM UNIMODAL COMBINATIONS OF FEATURES (TASK 2)

Feature	Female Partner				Male Partner			
	UA	Sensitivity	Specificity	AUC	UA	Sensitivity	Specificity	AUC
Self-Reported MQI	58.3	41.6	74.9	0.52	67.9**	52.4	83.5	0.60
Context	57.5	26.5	88.4	0.45	61.4	40.4	82.4	0.49
Interaction	55.7	43.4	68.1	0.45	69.5*	48.5	90.5	0.45
EDA	52.4	32.1	72.6	0.43	72.1**	54.5	89.6	0.58
ECG	61.1*	36.5	85.7	0.54	58.6	29.8	87.4	0.48
Linguistic	64.4**	52.9	76.0	0.58	58.7	41.8	75.5	0.54
Psychological	65.0**	54.2	75.8	0.53	62.5*	45.5	79.5	0.50
Personal	61.6*	36.2	86.9	0.45	70.0**	52.3	87.7	0.59
Paralinguistic	66.1**	56.6	75.6	0.58	61.3*	31.6	91.0	0.40
Audio	56.9	30.9	83.0	0.46	63.2*	39.9	86.5	0.45

Note. Task 2 presents percentages of UA, sensitivity, specificity, and AUC for the unimodal feature groups. UA = unweighted accuracy, AUC = area under the curve, MQI = mood and quality of interactions, EDA = electrodermal activity, ECG = electrocardiogram, * $p < 0.05$, ** $p < 0.01$ (UA significantly higher than 50% chance)

TABLE 3
CONFLICT CLASSIFICATION FROM MULTIMODAL COMBINATIONS (TASK 3) AND COMPARISONS WITH SELF-REPORTED MQI (TASK 4)

Feature Combination	Task 3				Task 4						
	UA	Sensitivity	Specificity	AUC	Stressed	Happy	Sad	Nervous	Angry	Close	Total
Female Partner											
EDA, Audio, Interaction, Context	78.2**	60.7	95.7	0.67	2.72*	1.85	2.05*	2.97**	1.57	2.63*	2.91**
ECG, Paralinguistic, Interaction, Context	76.3**	67.9	84.7	0.69	2.66*	1.75	1.95	2.93**	1.45	2.58*	2.85**
EDA, ECG, EDA Synchrony, Linguistic	75.5**	60.9	90.1	0.60	2.35*	1.51	1.69	2.68*	1.24	2.41*	2.55*
EDA, ECG, Paralinguistic, Personal, Interaction, Context	74.7**	61.8	87.6	0.73	2.31*	1.39	1.61	2.51*	1.09	2.31*	2.48*
EDA, EDA Synchrony, Linguistic, Personal	74.2**	61.9	86.5	0.58	2.20*	1.25	1.46	2.51*	0.94	2.14*	2.40*
Self-Reported MQI, EDA, Psychological, Personal, Acoustic, Interaction, Context	79.6**	73.5	85.7	0.79	2.88**	1.89	2.12*	3.20**	1.59	2.90**	3.09**
Self-Reported MQI, ECG, Psychological, Paralinguistic, Interaction, Context	79.4**	68.4	90.4	0.64	2.68*	1.74	1.95	2.97**	1.44	2.63*	2.88**
Male Partner											
EDA Synchrony, Psychological, Paralinguistic, Interaction, Context	79.3**	62.5	96.1	0.52	2.50*	2.27*	2.41*	3.10**	1.13	2.32*	1.08
EDA Synchrony, Psychological, Paralinguistic, Personal, Interaction, Context	78.3**	58.3	98.3	0.37	2.45*	2.16*	2.32*	2.97**	1.13	2.24*	0.96
Psychological, Personal, Interaction, Context	78**	60.5	95.5	0.64	2.19*	1.99	2.12*	2.76*	0.96	2.03	0.91
EDA Synchrony, Psychological, Personal, Interaction, Context	77.1*	58.4	95.8	0.52	2.14*	1.93	2.07	2.66*	0.92	1.99	0.85
EDA, Acoustic, Interaction, Context	76.9*	62.5	91.3	0.47	1.97	1.75	1.88	2.45*	0.81	1.80	0.73
Self-Reported MQI, Personal, Interaction, Context	86.8**	82.1	91.5	0.59	4.20**	3.83**	4.02**	4.82**	2.51*	3.92**	2.19*

Self-Reported MQI, EDA Synchrony, Paralinguistic, Personal, Acoustic, Interaction, Context	86.3**	80.1	92.5	0.64	3.69**	3.28**	3.48**	4.32**	1.92	3.37**	1.64
--	--------	------	------	------	--------	--------	--------	--------	------	--------	------

Note. Task 3 presents percentages of UA, sensitivity, specificity, and AUC for multimodal feature groups. Task 4 presents *t*-values for comparisons between the multimodal feature groups and self-reported MQI. UA = unweighted accuracy, AUC = area under the curve, total = combined self-reported MQI data, MQI = mood and quality of interactions, EDA = electrodermal activity, ECG = electrocardiogram, * $p < 0.05$, ** $p < 0.01$ (UA significantly higher than 50% chance).

TABLE 4
 CONFUSION MATRICES OF CONFLICT CLASSIFICATION FOR THE BEST PERFORMING MULTIMODAL COMBINATIONS (TASK 3)

		Female Partner		Male Partner	
		EDA, Audio, Interaction, Context		EDA Synchrony, Psychological, Paralinguistic, Interaction, Context	
		Predicted Conflict	Predicted No Conflict	Predicted Conflict	Predicted No Conflict
True Conflict		17	11	20	12
True No Conflict		6	134	5	125
		Self-Reported, Personal, Interaction, Context		Self-Reported, EDA, Psychological, Personal, Audio, Interaction, Context	
		Predicted Conflict	Predicted No Conflict	Predicted Conflict	Predicted No Conflict
True Conflict		22	8	23	5
True No Conflict		20	120	13	139

Note. Confusion matrices for classifying cases as conflict versus no conflict for the best performing multimodal combination with and without self-reported mood and quality of interactions.