Towards Real-Time Multimodal Emotion Recognition among Couples

George Boateng
gboateng@ethz.ch
ETH Zürich
Zurich, Switzerland

ABSTRACT

Researchers are interested in understanding the emotions of couples as it relates to relationship quality and dyadic management of chronic diseases. Currently, the process of assessing emotions is manual, time-intensive, and costly. Despite the existence of works on emotion recognition among couples, there exists no ubiquitous system that recognizes the emotions of couples in everyday life while addressing the complexity of dyadic interactions such as turn-taking in couples' conversations. In this work, we seek to develop a smartwatch-based system that leverages multimodal sensor data to recognize each partner's emotions in daily life. We are collecting data from couples in the lab and in the field and we plan to use the data to develop multimodal machine learning models for emotion recognition. Then, we plan to implement the best models in a smartwatch app and evaluate its performance in real-time and everyday life through another field study. Such a system could enable research both in the lab (e.g. couple therapy) or in daily life (assessment of chronic disease management or relationship quality) and enable interventions to improve the emotional well-being, relationship quality, and chronic disease management of couples.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Emotion Recognition; Multimodal Fusion; Couples; Smartwatches; Machine Learning; Deep Learning; Transfer Learning

ACM Reference Format:
George Boateng. 2020. Towards Real-Time Multimodal Emotion Recognition among Couples. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3382507.3421154

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3421154

1 INTRODUCTION

Romantic relationships have powerful effects on people's mental and physical health (see e.g. [68] for an overview). For instance, conflicts and negative qualities of one's intimate relationship are associated prospectively with morbidity and mortality [50]. Researchers are working towards understanding the emotional processes that take place in intimate relationships as underlying mechanisms for this relationship-health link (e.g. [33, 77]). Also, researchers are interested in assessing couples' emotions as they are affected in couples' dyadic management of chronic diseases [51] and they are predictors of relationship quality [37]. For example, spousal support in chronic disease management has been shown to have positive or negative effects on emotional well-being [15, 42, 63]. Also, emotions that couples experience during a conflict predict if these couples stay together in the long-term (for an overview, see [37]), and couples heading for break-up show more negative emotions and less positive emotions than happy couples, and are stuck in certain emotional patterns [19, 36].

However, assessing these emotions in couples is challenging. Two approaches are used for emotion assessment: self-report and observer reports. For self-reports, couples are asked to have emotionally charged conversation that is videotaped (e.g. in the lab) and then afterward, they provide emotion ratings while watching the videos [67]. These ratings could be biased and may not reflect the partner's actual emotion. In the case of daily life, couples are periodically asked to complete self-reports such as the PANAS [82] which can be obtrusive and impractical for continuous emotion assessment. For observers' reports, people are trained to watch the video recordings (e.g. in the case of lab data) and use a coding scheme to rate the interaction on specific emotional behaviors (e.g. SPAFF [24]). Such coding is also done for example, for audio data collected from couples' daily life interactions [66]. This manual coding process is costly and time-consuming as multiple coders need to be trained for this task [46] and suffers from inter-rater reliability issues [39, 54]. Automated emotion recognition could address these limitations, and therefore advance the field in important ways [58]. Yet, there exists no ubiquitous system that recognizes the emotions of couples in everyday life while addressing the complexity of dyadic interactions such as turn-taking in couples' conversations.

Smartwatches have been used for mood recognition of individuals [16] and they could be leveraged for emotion recognition among couples. Several features of smartwatches make them uniquely positioned for this task. Firstly, they are mostly with the wearer as opposed to a smartphone which could be in various places like the pocket, bag, and just not in proximity with the user. Additionally, commercial smartwatches could be used to collect a wide variety of sensor data that have been used for emotion recognition in the

past: audio [73], heart rate, accelerometer and gyroscope (for gestures) [72], and ambient light (to detect the context of couples). Our past work leveraged smartwatches for behavior recognition: e.g. tracking stress [9] and physical activity [7, 8]. Multimodal fusion of these sensor data could produce better recognition results [28, 62]. Finally and importantly, smartwatches could be leveraged in novel ways to capture the dyadic interactions of partners as we have done in our previous work (e.g. triggering data collection when partners are close and speaking) [12].

This research work seeks to develop a smartwatch-based system that leverages multimodal sensor data to recognize each partner's emotions in daily life. Such a system could enable research both in the lab (e.g. couple therapy) or in daily life (assessment of chronic disease management or relationship quality) and enable interventions to improve the emotional well-being, relationship quality and chronic disease management of couples. Towards this end, we seek to answer the following research questions (RQs).

RQ1: *How accurately can emotions be recognized using multimodal real-world sensor data from couples?* There are several challenges to address such as the kind of sensor data that should be collected, how the data should be fused together, what features to extract, what machine learning and deep learning approaches to use, how to evaluate the models, among others.

RQ2: *How accurately can the emotions of couples be detected in real-time in everyday life?* There are several challenges to address such as how well the algorithm will work on unseen couples across different cultures, when certain sensor data such as voice is not available, how to ensure that there is little latency in prediction, whether to do the prediction on a remote server considering various privacy issues or on-device, which will imply the machine learning model will need to be compact, potentially reducing the prediction accuracy, among others.

In the rest of this paper, we discuss background and related work in Section 2, methodology in Section 3, experiments and evaluation approach in Section 4 and results and contribution in Section 5.

2 BACKGROUND AND RELATED WORK

2.1 Emotion Models

There are mainly two models of emotions used in the literature in emotion recognition: categorical and dimensional. Categorical emotions are based on the six basic emotions proposed by Ekman: happiness, sadness, fear, anger, disgust, and surprise [30]. Dimensional approaches mainly use two dimensions: valence (pleasure) and arousal which are based on Russell's circumplex model of emotions [70]. Valence refers to how negative to positive the person feels and arousal refers to how sleepy to active a person feels. Using these two dimensions, several categorical emotions can be placed and grouped into the four quadrants: high arousal and negative valence (e.g. stress), low arousal and negative valence (e.g. depressed), low arousal and positive valence (e.g. relaxed) and high arousal and positive valence (e.g. excitement).

2.2 Multimodal Emotion Recognition

Multimodal fusion entails combining data collected from various modalities and leverages the idea that data contained in different modalities could provide a better understanding of a certain context. Various works have employed multimodal fusion approaches for emotion recognition and they have been shown to give better results than unimodal approaches [28, 62]. There are two main fusion approaches: fusion at the feature level (early fusion) i.e., combine features from different data modalities, for example, through concatenation and feeding them into the same machine learning algorithm or at the decision level (late fusion) i.e., have a different algorithm for each data modality and then combine the individual algorithm predictions using, for example, majority voting. Additional approaches include some hybrid of early and late fusion [83] and model-level fusion which leverages interactions between different modalities at the model level e.g [41].

2.3 Couple Emotion Recognition

Several emotion-recognition works on couple dyads have used data collected from individuals acting out dyadic interactions either using a script or engaging in spontaneous sessions [17, 18, 53, 55]. A lot of emotion recognition works use such data sets [62]. The emotions are later rated by others amidst several challenges [54] and do not necessarily reflect the subjective emotions of the individuals. Additionally, these algorithms are likely to perform poorly on naturalistic data [28].

On the other hand, there are a number of works on automatically detecting the emotions, and behavior of real couples. Most of this work with real couples has been done primarily by the Signal Analysis and Interpretation Laboratory (SAIL) at the University of South California with the first set of works published in 2010 [4, 47]. These works have ranged from the recognition of various behaviors of couples such as level of blame [4, 5], conflict [78], suicidal risk [23] to emotions [6, 22, 48] and led to the creation of the Behavioral Signal Processing (BSP) domain [35, 56]. Few works done by researchers outside this research lab include [25, 29].

For the works focused on recognition of emotions, they used mostly acoustic data [4, 6, 25, 29, 47–49, 84], and others have used lexical data [20, 21, 79, 80] and few have used visual data [85] and multimodal data such as speech and lexical data [22, 45, 81]. A range of algorithms have been used such as linear discriminant analysis (LDA), support vector machine (SVM), logistic regression, hidden Markov models, deep neural networks. Evaluations have mostly been done with leave-one-couple-out cross-validation which is a robust evaluation approach for this task. Also, several works trained gender-specific models as there is a gender difference in some modalities like speech.

Despite the contributions of these works, there are still significant gaps remaining. All these works used emotion labels from external raters rather than the couples and hence do not reflect the subjective emotions of the couples. Only two modalities have been used with several modalities such as physiological data and hand gestures unexplored. Most have used data from the lab and none of these models have been tested for real-time emotion recognition among couples in daily life. Hence, to date, there exists no ubiquitous system that recognizes the emotions of couples in real-time in daily life while addressing noisiness of real-world data and the complexity of dyadic interactions such as turn-taking in couples' conversations.

3 METHODOLOGY

To answer our research questions, our plan is to implement the following approach:

(1) Develop mobile and wearable apps and collect multimodal sensor and self-report data about emotions from couples in the lab and everyday life
(2) Develop multimodal emotion-recognition machine-learning models using the collected data
(3) Implement the model on a smartwatch to perform real-time recognition of couples' emotions

3.1 Data Collection

We plan to use two datasets of emotion data from couples in Belgium (collected in the lab) and Switzerland (collection from field and lab ongoing).

3.1.1 Study 1: Dyadic Interaction Study. A Dyadic Interaction lab study was conducted in Leuven, Belgium with 101 Dutch-speaking, couples. These couples were first asked to have a 10-minute conversation about a negative topic (a characteristic of their partner that annoys them the most), followed by a 10-minute conversation about a positive topic (a characteristic of their partner that they value the most) [26, 74–76]. During both conversations, couples were asked to wrap up the conversation after 8 minutes. For the negative topic, they were also asked to end on good terms. After each conversation, each partner completed self-reports on various categorical emotion labels such as anger, sadness, anxiety, relaxation, happiness, etc. on a 7-point Likert scale ranging from strongly disagree (1) to strongly agree (7). Additionally, each partner watched the video recording of the conversation separately on a computer and rated his or her emotion on a moment-by-moment basis by continuously adjusting a joystick to the left (very negative) and the right (very positive), so that it closely matched their feelings, resulting in valence scores on a continuous scale from -1 to 1 [38, 69]. Additionally, each partner reported how they felt after the interaction and how they thought their partner felt, using the Affect Grid questionnaire [71]. The Affect Grid captures the valence and arousal dimensions of Russell's circumplex model of emotions [70]. Subjects had to place an 'x' on any square on the Affect Grid corresponding to their feelings about each conversation, which translates to a value of between 0 and 8 each for pleasure and arousal.

3.1.2 Study 2: DyMand Study. We are currently running a Dyadic Management of Diabetes (DyMand) field and lab study in Switzerland with German-speaking couples in which one partner has type 2 diabetes. We plan to collect data from 180 couples (N=180; n=360) but we have collected data from ten (10) couples so far [51]. We collect data from the field for 7 days and also in the lab after the couples return.

For the field study, each partner is given a smartwatch and smartphone running the DyMand system, a novel open-source mobile and wearable system that we developed for ambulatory assessment of couples' chronic disease management [12]. The DyMand system triggers the collection of sensor and self-report data for 5 minutes each hour during the hours that subjects pick. We collect the following sensor data from the smartwatch: audio, heart rate, accelerometer, gyroscope, Bluetooth low energy (BLE) signal strength between watches, and ambient light. After the sensor data collection, a self-report is triggered on the smartphone that asks about emotions over the last 5 minutes using the Affective Slider, a digital affect measuring tool which measures which assesses the valence and arousal dimensions of their emotions [3]. We also record a 3-second video of their facial expression while they complete the self-report on the smartphone. Additionally, at the end of the day, we trigger the Affective Slider, and also a short form of the PANAS self-report [82] for the couples to report their emotions over the whole day.

Our hypothesis is that we are likely to collect high-quality sensor and self-report emotion data during times that the partners are interacting. Hence, rather than trigger data collection at some random times in the hour which is the standard approach [52, 66], we use a novel method entailing triggering data collection after we detect that the partners are close and speaking. We trigger sensor data collection when the partners are close and speaking in two steps. First, we determine closeness using the BLE signal strength between the smartwatches. We check if the signal strength is within a certain threshold, which corresponds to a distance estimate [13]. Then, we determine if the partners are speaking by using a voice activity detection (VAD) machine learning model that classifies speech versus non-speech, which we developed and implemented to run in real-time on the smartwatch [12]. In the case in which this condition is not met in the hour, we do a backup recording in the last 15 minutes of the hour. There are significant ethical and privacy concerns of such a system and study which we address in our previous works [11, 13].

For the lab study, the couple is asked to discuss an illness management–related concern that is causing them considerable distress for a 10-minute period. The session is videotaped and additionally, each partner wears a smartwatch running the DyMand app as it collects various sensor data: audio, heart rate, accelerometer, gyroscope, and ambient light. After the session, each partner completes a self-report on a smartphone about their emotions over the last 10 min of the discussion using the Affective Slider [3]. Also, the smartphone takes a 3-second video of their facial expression while they complete the self-report.

3.2 Data Preprocessing

We will preprocess the sensor data into a form for easy data analysis. For the audio data, we will remove nonvocal segments for the field data (e.g. silence and noise portions), filter, downsample, and reduce the background noise. We will perform speaker diarization to annotate the segments of the audio corresponding to the speech of each partner. We will also annotate any segments of the audio corresponding to various nonverbal vocalizations such as laughs, sighs, and also background context especially for the field data e.g. TV, audio, indoors, outdoors, etc. Additionally, we will also transcribe the audio in order to use the content of the speech. The speaker diarization, annotation, and transcriptions are done manually to ensure high data quality. Then, we will develop automated tools to do same for the real-time recognition. Audio samples that are found to be too noisy to be useful will not be used for data analysis. Other sensor data such as accelerometer and gyroscope data will be filtered and downsampled. Heart rate data will be processed to

remove samples that were collected when there was no or poor contact with the skin since the smartwatch provides that data.

3.3 Feature Extraction

We plan to extract features using feature engineering and transfer learning approaches.

3.3.1 Feature Engineering. For audio, we plan to extract various acoustic time-domain features such as pitch, speech rate, etc, and also frequency domain features such as spectral energy, fundamental frequency, etc. We plan to use the OpenSMILE toolkit [32] to extract 88 feature sets that have been shown to be a minimal feature set that works well for the task acoustic emotion recognition [31]. We also plan to use the presence or absence of various nonverbal vocalizations as features as they have been shown to be discriminative for emotion recognition [43]. We will also extract spectrograms of the audio data to use as alternate features. For accelerometer and gyroscope data, heart rate, and ambient light, we will extract various statistics like mean, median, and percentiles over various durations.

3.3.2 Transfer Learning. Transfer learning is an approach used to circumvent the need to develop hand-crafted features usually done in traditional machine learning approaches and also deal with small labeled datasets. Transfer learning entails using a pre-trained model on a different but related task ([34]). This process entails using the model for feature extraction or fine-tuning in which the whole model or later layers are retrained. Transfer learning has shown success in various fields such as computer vision ([44, 60]), speech processing ([59]), and natural language processing ([40, 65]). Transfer learning has also been used in emotion recognition tasks ([34, 57]). We plan to use a pretrained acoustic CNN model such as YAMNet model [2] to extract acoustic features and a pretrained Transformer language model — Bidirectional Encoder Representations from Transformers (BERT) [27] such as the German BERT [1] and Sentence BERT [64] — to extract linguistic features, which we've used in a previous work ([10]).

3.4 Data Analysis

To analyze the data, we plan to use various machine learning and deep learning algorithms as well as explore unimodal and multimodal analysis of the data.

3.4.1 Machine Learning and Deep Learning Algorithms. We plan to explore using traditional machine learning algorithms such as random forest, support vector machines and for deep learning, we will explore using convolutional neural networks and recurrent neural networks such as Long Short-Term Memory (LSTM) — with and without attention — bidirectional LSTM (BLSTM), together with handcrafted features ([61]). We will also explore using the raw signal in an end-to-end approach leveraging 1D CNNs and LSTMs.

3.4.2 Unimodal and Multimodal Analysis. We plan to evaluate models separately for each modality. Additionally, we will explore multimodal fusion using different combinations of modalities at the feature level, the decision level, model level or some hybrid approach.

4 EXPERIMENTS AND EVALUATION

We will train models separately for males and females to perform binary classification of valence and arousal as has been done in previous works. We plan to perform an evaluation with leave-one-couple-out cross-validation similar to previous works such as [22] with the metrics confusion matrix and balanced accuracy since the data is likely to be imbalanced which is characteristic of real-world data. We plan to evaluate performance when various modalities are left out. We will also compare the performance of using features extracted with data annotated manually vs automatically.

We will then pick the best performing algorithm and then optimize it to run in real-time and develop a smartwatch app using the model. We will then run another user study to evaluate the performance of the model in real-time in everyday life. The system will trigger data collection when an interaction is detected like the previously described Study 2. We will additionally for evaluation purposes trigger at some random times also. After 5 minutes of data collection, subjects will be asked to respond to self-reports about emotions using the Affective Slider. These self-reports will be compared with the system's predictions to evaluate its performance.

5 RESULTS AND CONTRIBUTIONS

We have developed the DyMand mobile and wearable system along with a smartwatch-based VAD system for sensor and self-report data collection for study 2 [12, 13]. We have performed preliminary analysis using the data from study 1 on speech emotion recognition among couples. We performed an evaluation of the segments of an audio conversation that best predicts the end-of-conversation emotions of couples. We leveraged the peak-end rule, and a used transfer learning approach to extract features from (1) the audio segments with the most extreme positive and negative ratings, and (2) the ending of the audio. We used a pre-trained CNN to extract these acoustic features and a linear SVM to perform binary classification of the valence of partners. Our results showed that the segments from the peak produce the best results for recognizing the emotions of female partners with 74.3% balanced accuracy and better than chance and a human baseline [14].

There are two main potential contributions of my research work if the research questions are adequately answered.

(1) A novel machine learning method for emotion recognition using multimodal real-world smartwatch data from couples
(2) A novel smartwatch app for real-time emotion recognition among couples in everyday life

These contributions would provide an easy assessment of emotions and could enable research both in the lab (e.g. couple therapy) and in daily life (e.g. assessment of chronic disease management or relationship quality) and enable interventions to improve the emotional well-being, relationship quality and chronic disease management of couples.

ACKNOWLEDGMENTS

I'm grateful to my research advisors, Prof. Dr. Tobias Kowatsch, and Prof. Dr. Elgar Fleisch. Also, I thank my research collaborators and lab colleagues for their contributions and assistance in this research work. This research is partially funded by the Swiss National Science Foundation (CR12I1_166348/1).

REFERENCES

[1] [n.d.]. Open Sourcing German BERT. https://deepset.ai/german-bert. Accessed: 2020-05-1.

[2] [n.d.]. YAMNet. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet.

[3] Alberto Betella and Paul FMJ Verschure. 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one* 11, 2 (2016), e0148037.

[4] Matthew Black, Athanasios Katsamanis, Chi-Chun Lee, Adam C Lammert, Brian R Baucom, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2010. Automatic classification of married couples' behavior using audio features. In *Eleventh annual conference of the international speech communication association.*

[5] Matthew P Black, Panayiotis G Georgiou, Athanasios Katsamanis, Brian R Baucom, and Shrikanth Narayanan. 2011. "You made me do it": Classification of Blame in Married Couples' Interactions by Fusing Automatically Derived Speech and Language Information. In *Twelfth Annual Conference of the International Speech Communication Association.*

[6] Matthew P Black, Athanasios Katsamanis, Brian R Baucom, Chi-Chun Lee, Adam C Lammert, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2013. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech communication* 55, 1 (2013), 1–21.

[7] George Boateng, John A Batsis, Ryan Halter, and David Kotz. 2017. ActivityAware: an app for real-time daily activity level monitoring on the amulet wrist-worn device. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 431–435.

[8] George Boateng, John A Batsis, Patrick Proctor, Ryan Halter, and David Kotz. 2018. GeriActive: Wearable app for monitoring and encouraging physical activity among older adults. In *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 46–49.

[9] George Boateng and David Kotz. 2016. StressAware: An app for real-time stress monitoring on the amulet wearable platform. In *2016 IEEE MIT Undergraduate Research Technology Conference (URTC)*. IEEE, 1–4.

[10] George Boateng and Tobias Kowatsch. 2020. Speech Emotion Recognition among Elderly Individuals using Transfer Learning. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.*

[11] George Boateng, Janina Lüscher, Urte Scholz, and Tobias Kowatsch. 2020. Emotion Capture among Real Couples in Everyday Life. In *1st Momentary Emotion Elicitation & Capture workshop (MEEC 2020), co-located with the ACM CHI Conference on Human Factors in Computing Systems.*

[12] George Boateng, Prabhakaran Santhanam, Janina Lüscher, Urte Scholz, and Tobias Kowatsch. 2019. Poster: DyMand–An Open-Source Mobile and Wearable System for Assessing Couples' Dyadic Management of Chronic Diseases. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–3.

[13] George Boateng, Prabhakaran Santhanam, Janina Lüscher, Urte Scholz, and Tobias Kowatsch. 2019. VADLite: an open-source lightweight system for real-time voice activity detection on smartwatches. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 902–906.

[14] George Boateng, Laura Sels, Peter Kuppens, Peter Hilpert, Urte Scholz, and Tobias Kowatsch. 2020. Speech Emotion Recognition among Couples using the Peak-End Rule and Transfer Learning. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.*

[15] Niall Bolger and David Amarel. 2007. Effects of social support visibility on adjustment to stress: Experimental evidence. *Journal of personality and social psychology* 92, 3 (2007), 458.

[16] Pascal Budner, Joscha Eirich, and Peter A Gloor. 2017. " Making you happy makes me happy"-Measuring Individual Mood with Smartwatches. *arXiv preprint arXiv:1711.06134* (2017).

[17] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[18] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing* 8, 1 (2016), 67–80.

[19] Laura L Carstensen, John M Gottman, and Robert W Levenson. 1995. Emotional behavior in long-term marriage. *Psychology and aging* 10, 1 (1995), 140.

[20] Sandeep Nallan Chakravarthula, Brian Baucom, and Panayiotis Georgiou. 2018. Modeling Interpersonal Influence of Verbal Behavior in Couples Therapy Dyadic Interactions. *arXiv preprint arXiv:1805.09436* (2018).

[21] Sandeep Nallan Chakravarthula, Rahul Gupta, Brian Baucom, and Panayiotis Georgiou. 2015. A language-based generative model framework for behavioral analysis of couples' therapy. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2090–2094.

[22] Sandeep Nallan Chakravarthula, Haoqi Li, Shao-Yen Tseng, Maija Reblin, and Panayiotis Georgiou. 2019. Predicting Behavior in Cancer-Afflicted Patient and Spouse Interactions Using Speech and Language. *Proc. Interspeech 2019* (2019), 3073–3077.

[23] Sandeep Nallan Chakravarthula, Md Nasir, Shao-Yen Tseng, Haoqi Li, Tae Jin Park, Brian Baucom, Craig J Bryan, Shrikanth Narayanan, and Panayiotis Georgiou. 2020. Automatic Prediction of Suicidal Risk in Military Couples Using Multimodal Interaction Cues from Couples Conversations. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6539–6543.

[24] James A Coan and John M Gottman. 2007. The specific affect coding system (SPAFF). *Handbook of emotion elicitation and assessment* (2007), 267–285.

[25] Colleen E Crangle, Rui Wang, Marcos Perreau-Guimaraes, Michelle U Nguyen, Duc T Nguyen, and Patrick Suppes. 2019. Machine learning for the recognition of emotion in the speech of couples in psychotherapy using the Stanford Suppes Brain Lab Psychotherapy Dataset. *arXiv preprint arXiv:1901.04110* (2019).

[26] Egon Dejonckheere, Merijn Mestdagh, Marlies Houben, Isa Rutten, Laura Sels, Peter Kuppens, and Francis Tuerlinckx. 2019. Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature human behaviour* 3, 5 (2019), 478–491.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[28] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 1–36.

[29] Harishchandra Dubey, Matthias R Mehl, and Kunal Mankodiya. 2016. Bigear: Inferring the ambient and emotional correlates from smartphone-based acoustic big data. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 78–83.

[30] Paul Ekman and Dacher Keltner. 1997. Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture* (1997), 27–46.

[31] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[32] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[33] Allison K Farrell, Ledina Imami, Sarah CE Stanton, and Richard B Slatcher. 2018. Affective processes as mediators of links between close relationships and physical health. *Social and Personality Psychology Compass* 12, 7 (2018), e12408.

[34] Kexin Feng and Theodora Chaspari. 2020. A Review of Generalizable Transfer Learning in Automatic Emotion Recognition. *Frontiers in Computer Science* 2 (2020), 9.

[35] Panayiotis G Georgiou, Matthew P Black, and Shrikanth S Narayanan. 2011. Behavioral signal processing for understanding (distressed) dyadic interactions: some recent developments. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. 7–12.

[36] John Mordechai Gottman. 2005. *The mathematics of marriage: Dynamic nonlinear models*. MIT Press.

[37] John Mordechai Gottman. 2014. *What predicts divorce?: The relationship between marital processes and marital outcomes*. Psychology Press.

[38] John M Gottman and Robert W Levenson. 1985. A valid procedure for obtaining self-report of affect in marital interaction. *Journal of consulting and clinical psychology* 53, 2 (1985), 151.

[39] Richard E Heyman. 2001. Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations. *Psychological assessment* 13, 1 (2001), 5.

[40] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

[41] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. 2020. Multimodal Transformer Fusion for Continuous Emotion Recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3507–3511.

[42] Masumi Iida, Mary Ann Parris Stephens, Karen S Rook, Melissa M Franks, and James K Salem. 2010. When the going gets tough, does support get going? Determinants of spousal support provision to type 2 diabetic patients. *Personality and Social Psychology Bulletin* 36, 6 (2010), 780–791.

[43] Roza Kamiloglu, George Boateng, Alisa Balabanova, Chuting Cao, and Disa Sauter. 2020. Many ways to sound good: Superior decoding of positive emotions from nonverbal vocalisations compared to speech prosody. *PsyArXiv* (2020).

[44] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional

neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[45] Athanasios Katsamanis, James Gibson, Matthew P Black, and Shrikanth S Narayanan. 2011. Multiple instance learning for classification of human behavior observations. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 145–154.

[46] Patricia K Kerig and Donald H Baucom. 2004. *Couple observational coding systems*. Taylor & Francis.

[47] Chi-Chun Lee, Matthew Black, Athanasios Katsamanis, Adam C Lammert, Brian R Baucom, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Eleventh Annual Conference of the International Speech Communication Association*.

[48] Chi-Chun Lee, Athanasios Katsamanis, Matthew P Black, Brian R Baucom, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2014. Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions. *Computer Speech & Language* 28, 2 (2014), 518–539.

[49] Haoqi Li, Brian Baucom, and Panayiotis Georgiou. 2016. Sparsely connected and disjointly trained deep neural networks for low resource behavioral annotation: Acoustic classification in couples' therapy. *arXiv preprint arXiv:1606.04518* (2016).

[50] Timothy J Loving and Richard B Slatcher. 2013. Romantic relationships and health. *The Oxford handbook of close relationships* (2013), 617–637.

[51] Janina Lüscher, Tobias Kowatsch, George Boateng, Prabhakaran Santhanam, Guy Bodenmann, and Urte Scholz. 2019. Social Support and Common Dyadic Coping in Couples' Dyadic Management of Type II Diabetes: Protocol for an Ambulatory Assessment Application. *JMIR research protocols* 8, 10 (2019), e13685.

[52] Matthias R Mehl, Megan L Robbins, and Fenne große Deters. 2012. Naturalistic observation of health-relevant social processes: The Electronically Activated Recorder (EAR) methodology in psychosomatics. *Psychosomatic Medicine* 74, 4 (2012), 410.

[53] Angeliki Metallinou, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, Shrikanth Narayanan, et al. 2010. The USC CreativeIT database: A multimodal database of theatrical improvisation. *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality* (2010), 55.

[54] Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–8.

[55] Philipp M Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 663–669.

[56] Shrikanth Narayanan and Panayiotis G Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proc. IEEE* 101, 5 (2013), 1203–1233.

[57] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 443–449.

[58] Sally Olderbak, Andrea Hildebrandt, Thomas Pinkpank, Werner Sommer, and Oliver Wilhelm. 2014. Psychometric challenges and proposed solutions when scoring facial emotion expression codes. *Behavior Research Methods* 46, 4 (2014), 992–1006.

[59] Charles C. Onu, Jonathan Lebensold, William L. Hamilton, and Doina Precup. 2019. Neural Transfer Learning for Cry-Based Diagnosis of Perinatal Asphyxia. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, Gernot Kubin and Zdravko Kacic (Eds.). ISCA, 3053–3057. https://doi.org/10.21437/Interspeech.2019-2340

[60] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1717–1724.

[61] Sandeep Kumar Pandey, HS Shekhawat, and SRM Prasanna. 2019. Deep Learning Techniques for Speech Emotion Recognition: A Review. In *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 1–6.

[62] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[63] Gabriele Prati and Luca Pietrantoni. 2010. The relation of perceived and received social support to mental health among first responders: a meta-analytic review. *Journal of Community Psychology* 38, 3 (2010), 403–417.

[64] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[65] Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpiDEDIS at GermEval 2019: Offensive Language Identification using a German BERT model. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). Erlangen, Germany: German Society for Computational Linguistics & Language Technology*. 403–408.

[66] Megan L Robbins, Ana María López, Karen L Weihs, and Matthias R Mehl. 2014. Cancer conversations in context: naturalistic observation of couples coping with breast cancer. *Journal of Family Psychology* 28, 3 (2014), 380.

[67] Nicole A Roberts, Jeanne L Tsai, and James A Coan. 2007. Emotion elicitation using dyadic interaction tasks. *Handbook of emotion elicitation and assessment* (2007), 106–123.

[68] Theodore F Robles, Richard B Slatcher, Joseph M Trombello, and Meghan M McGinn. 2014. Marital quality and health: A meta-analytic review. *Psychological bulletin* 140, 1 (2014), 140.

[69] Anna Marie Ruef and Robert W Levenson. 2007. Continuous measurement of emotion. *Handbook of emotion elicitation and assessment* (2007), 286–297.

[70] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[71] James A Russell, Anna Weiss, and Gerald A Mendelsohn. 1989. Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology* 57, 3 (1989), 493.

[72] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. 2019. Wearable-Based Affect Recognition—A Review. *Sensors* 19, 19 (2019), 4079.

[73] Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 5 (2018), 90–99.

[74] Laura Sels, Jed Cabrieto, Emily Butler, Harry Reis, Eva Ceulemans, and Peter Kuppens. 2019. The occurrence and correlates of emotional interdependence in romantic relationships. *Journal of personality and social psychology* (2019).

[75] Laura Sels, Eva Ceulemans, and Peter Kuppens. 2019. All's well that ends well? A test of the peak-end rule in couples' conflict discussions. *European Journal of Social Psychology* 49, 4 (2019), 794–806.

[76] Laura Sels, Yan Ruan, Peter Kuppens, Eva Ceulemans, and Harry Reis. 2020. Actual and perceived emotional similarity in couples' daily lives. *Social Psychological and Personality Science* 11, 2 (2020), 266–275.

[77] Timothy W Smith and Karen Weihs. 2019. Emotion, social relationships, and physical health: concepts, methods, and evidence for an integrative perspective. *Psychosomatic medicine* 81, 8 (2019), 681–693.

[78] Adela C Timmons, Theodora Chaspari, Sohyun C Han, Laura Perrone, Shrikanth S Narayanan, and Gayla Margolin. 2017. Using multimodal wearable technology to detect conflict among couples. *Computer* 50, 3 (2017), 50–59.

[79] Shao-Yen Tseng, Brian R Baucom, and Panayiotis G Georgiou. 2017. Approaching Human Performance in Behavior Estimation in Couples Therapy Using Deep Sentence Embeddings.. In *INTERSPEECH*. 3291–3295.

[80] Shao-Yen Tseng, Sandeep Nallan Chakravarthula, Brian R Baucom, and Panayiotis G Georgiou. 2016. Couples Behavior Modeling and Annotation Using Low-Resource LSTM Language Models.. In *INTERSPEECH*. 898–902.

[81] Shao-Yen Tseng, Haoqi Li, Brian Baucom, and Panayiotis Georgiou. 2018. " Honey, I Learned to Talk" Multimodal Fusion for Behavior Analysis. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 239–243.

[82] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[83] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28, 3 (2013), 46–53.

[84] Wei Xia, James Gibson, Bo Xiao, Brian Baucom, and Panayiotis G Georgiou. 2015. A dynamic model for behavioral analysis of couple interactions using acoustic features. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[85] Bo Xiao, Panayiotis Georgiou, Brian Baucom, and Shrikanth S Narayanan. 2015. Head motion modeling for human behavior analysis in dyadic interaction. *IEEE transactions on multimedia* 17, 7 (2015), 1107–1119.