Unfolding the External Behavior and Inner Affective State of Teammates through Ensemble Learning: Experimental Evidence from a Dyadic Team Corpus

Aggeliki Vlachostergiou[1], Mark Dennison[2], Catherine Neubauer[2], Stefan Scherer[3], Peter Khooshabeh[2], Andre Harrison[2]
[1]Image,Video and Multimedia Systems Lab, National Technical University of Athens, Athens, Greece
[2]US Army Research Lab, Los Angeles, CA, USA
[3]Institute of Creative Technologies, University of Southern California, Los Angeles, CA, USA
aggelikivl@image.ntua.gr, {mark.s.dennisson.ctr, catherine.e.neubauer2.ctr, peter.khooshabehadeh2.civ, andre.v.harrison2.civ}@mail.mil, scherer@ict.usc.edu

Abstract

The current study was motivated to understand the relationship between the external behavior and inner affective state of two team members ("instructor"-"defuser") during a demanding operational task (i.e., bomb defusion). In this study we assessed team member's verbal responses (i.e., length of duration) in relation to their external as well as internal affective cues. External behavioral cues include defuser's verbal expressions while inner cues are based on physiological signals. More specifically, we differentiate between "defusers'" physiological patterns occurring after the "instructor's" turns according to whether they belong to a short or a long turn-taking response interval. Based on the assumption that longer turn-taking behaviors are likely to be caused by demanding cognitive task events and/or stressful interactions, we hypothesize that inner mechanisms produced in these intense affective activity intervals will be reflected on defuser's physiology. A dyadic team corpus was used to examine the association between the "defusers" physiological signals following the "instructor's" questions to predict whether they occurred in a short or long turn-taking period of time. The results suggest that an association does exist between turn taking and inner affective state. Additionally, it was our goal to further unpack this association by creating diverse ensembles. As such, we studied various base learners and different ensemble sizes to determine the best approach towards building a stable diverse ensemble that generalizes well on the external and inner cues of individuals.
Keywords: Turn-takings, Physiological features, MEAP, Dyadic Team Corpus, Ensemble Learning

1. Introduction

Lack of emotional expressivity is one of the main deficits that characterizes periods of stress when team members perform highly cognitive cooperative tasks. In cases such as this, teammates may find it more difficult to express their conscious feelings and show different patterns in perceiving and conveying emotional information when working together to meet a common goal(Jones and George, 1998; Prati et al., 2003). In light of these observations, having a way to monitor the internal state of teammates within such contexts might provide us new insights with respect to the mechanisms of their interaction and affectivity.

During high workload and high stress tasks, the sympathetic nervous system is accountable for activating glands and organs that are responsible for defending the body from perceived threats. This activation is associated with changes in arousal that are further influenced by emotion, cognition or attention. Stress results in increased sympathetic activity and can be tracked for example through bodily reactions, such as an increase in heart rate, greater blood flow to extremities and an increase in the respiration rate etc. Thus, a combination of more than one physiological indicator would be considered a more sensitive measure of changes in stress and can be used to provide estimations of emotion, arousal and general cognition (McEwen, 2007).

In this paper, we shed light on the association between two team members' physiological states and their speech, measured via their conversational turn-taking duration. Team members in highly-demanding operational tasks do not often notice triggers that cause them to be emotionally and mentally stressed (Murphy, 1996; Stein, 2001). Thus, they might communicate with their teammates and express their emotions in ways that may not be noticed in observable audio-visual cues. For instance, one would expect that asking an individual to disarm a simulated bomb would result in high levels of (internal) stress. This inherent gap between teammates' external observable behavior and their inner affective state is not well understood and can be potentially bridged by monitoring their physiology. The duration of response utterances is also reported to be very important, as it can be indicative of conflicting mental and stress procedures (Raux and Eskenazi, 2009). Because physiological indicators reflect aspects of underlying mental states and specifically the amount of distress (El-Sheikh et al., 1989), we explore whether physiological signals of long and short response utterance durations exhibit different physiological patterns. To further capture and interpret this ongoing and evolving interplay, we examine the use of two ensemble learning strategies. We believe that the investigation of physiological changes during the response periods can provide a better understanding of a team dynamics.

2. Related Work

There has been a lot of research on dialogue dynamics as well as the relationship of stress to underlying physiology, but these fields have largely been separate in the literature. For example, work on turn-taking behavior in dialogue systems (Raux and Eskenazi, 2009) could benefit from an understand of interpersonal dynamics of physiological stress response during a cooperative task (Dennison et al., 2016). Links between turn-taking behavioral responses and physiology have been studied for assessing how adults' anger

levels affected children (El-Sheikh et al., 1989). Moreover, previous studies have shown the advantages of using ensemble learning in both unimodal (Schuller et al., 2005b; Scherer et al., 2008; Schels and Schwenker, 2010) and multimodal behavioral analysis(Glodek et al., 2011; Schels et al., 2012; Schuller et al., 2005a).

To the best of our knowledge, there is no experimental evidence of applying ensemble learning to study the link between external behavior of turn-taking responses (Section 4.) and inner affective states inferred from physiological signal indicators (Section 5.) of two teammates (instructor-defuser) trying to disarm a simulated bomb (Section 3.)(Neubauer et al., 2016). This work is an effort to unfold this association based on the experimental evidence from the Dyadic Team Corpus. Our results indicate that physiological patterns convey information about the defuser's inner state, because they differ according to the duration of turn-taking behavioral replies with respect to the instructor's turns (Section 6.2.). Finally, our results are further enhanced through ensemble learning methods, which outperform the individual base learners in most cases and interesting observations are discussed in the Section 7..

3. Corpus Description

The dyadic cooperative team corpus (Neubauer et al., 2016) employed a 2x2 between subjects design resulting in a total of 2 experimental conditions with 20-gender-matched pairs in the following two conditions: The Ice Breaker conversation (IB) condition which consisted of allowing teammates to garner familiarity through a series of "getting to know you" questions prior to the start of the task and the Control (CT) condition, where teammates simply began the task with no prior familiarity. The corpus consists of a series of simulated "bomb defusion" scenarios. In each scenario one team member served as the "defuser" and one team member served as the "instructor". The "instructor" was given a manual with instructions on how to diffuse the bomb. The "instructor" was told that it was their responsibility to provide information that would allow the defuser to successfully complete the task. After each separate task the team members switched roles (i.e., each team member was given the opportunity to be both the "defuser" and the "instructor" twice during the main task), which resulted in a total of 4 main tasks, each lasting an average of 5mins. For this work, we take into account only 10-gender-matched pairs (5 from each condition) and we examine only the case in which participant A is the instructor and participant B is the defuser, (i.e. we didn't examine the case of asking members to switch roles).

4. Turn-taking behavioral responses

One of the main indicators of an interactional speech episode is often called a "turn-taking" and is defined as the time duration between the end of someone's turn and the beginning of the other interlocutor's corresponding turn. Turn-taking responses may span from very short to very long, which may indicate shorter or longer emotional and stressful episodes. In a similar way, in our teammate corpus, longer turn taking behavioral responses provided valuable information about the defuser's perceived cognition and affective state, reflected their external observable as well as implicit inner affective states. We choose to investigate that type of interactional context between the two teammates, motivated by the fact that the instructor's behavior is more controllable, thus minimizing the effect of the instructor's variability on the defuser's behavior.

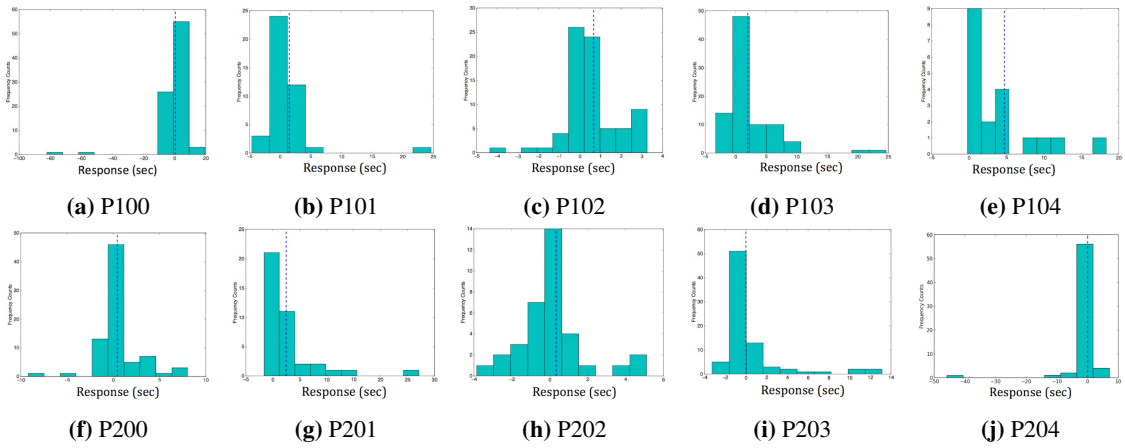
To further distinguish between short and long turn-taking behavioral responses (Figure 1) we draw a threshold at the 70th percentile of response values. This threshold was computed empirically after plotting the histograms of turn-taking behavioral response instances from the data of each defuser separately. Negative values of this measure mean that the defuser started talking before the instructor had finished the current turn. Phenomena such as overlapped speech and very short utterances are aligned with high levels of stress in highly-demanding operational tasks (Heldner and Edlund, 2010).

After carefully inspecting various turn-taking behavioral instances in our corpus, we came across a number of interesting tendencies. There were examples during the interactional context in which the instructor explained how the blue and red wires are connected. In that case, the defuser's reply is short (i.e., the defuser uses words such as ok/yes/no). Then, as a follow up, the instructor explained with more detail the process of bomb defusion to check whether the defuser is really following his instructions. In the first case, where the statement is simple and elicits low cognitive effort, a short reply occurred, while a long one occurred in the second case, where the defuser repeated the instructor's guidelines to confirm that he correctly understood the task. In this case, it was expected that the defuser was much more mentally alert.

5. Extraction of Physiological Features

A BIOPAC MP150, with a standard lead II electrode configuration was used to record electrocardiography (ECG), real-time changes in blood pressure, and impedance cardiography (ZKG). Continuous data were recorded for each participant throughout each task and analyzed offline. The raw time series for each task segmented into thirty second intervals relative to the end of each session, such that a few seconds from the beginning of the "bomb defusion" task were cut out. This was done because a minimum of 30 seconds of cardiovascular data are necessary for further analysis. The Moving Ensemble Average Program (MEAP)(Cieslak, 2017) was used to extract features from the data by computing an ensembled average over each epoch.

We extracted 24 cardiovascular features. Some of these features, presented in Table 1, included heart rate (HR), LVET (left ventral (systolic) ejection time), p_time, s_time, t_time, x_time, systole_time, pre-ejection period (PEP), ventricular contractility (VC), cardiac output (CO) and total peripheral resistance (TPR). PEP is the time from the onset of the heart muscle depolarization to the opening of the aortic valve. When PEP decreases, VC increases. VC has been shown to be related to task engagement (Newlin and Levenson, 1979; Richter and Gendolla, 2009; Spangler and Friedman, 2015; Seery, 2011). CO is the amount of blood pumped in liters per minute. TPR reflects vasodilation (more blood flow) and vasoconstriction (less blow



(a) P100 **(b)** P101 **(c)** P102 **(d)** P103 **(e)** P104

(f) P200 **(g)** P201 **(h)** P202 **(i)** P203 **(j)** P204

Figure 1: Frequency counts (y axis) for 10 defusers in IB and CT conditions and distributions of their responses (x axis) measured in seconds with respect to the instructor's turns. The vertical dashed blue line distinguishes between the short and long replies.

Participants	Condition	Selected Physiological Features
P100	IB	s_time, systole_time
P101	IB	hr
P102	IB	lvet,p_time, x_time
P103	IB	hr
P104	IB	t_time
P200	CT	diastole_time
P201	CT	hr
P202	CT	pep
P203	CT	hr
P204	CT	hr

Table 1: Most frequently selected Physiological Features for ten defusers, five from the IB and CT conditions respectively during the "bomb defusion" task. The features are labeled as follows: heart rate (hr), left ventral (systolic) ejection time (LVET) and pre-ejection period (pep).

flow), which are related to parasympathetic and sympathetic activity, respectively. Prior work has shown that TPR unambiguously increases when an individual is in a threat state and decreases in a challenge state, whereas CO either remains unchanged or decreases in a threat state and increases in a challenge state (Tomaka et al., 1997).

6. Experiments

The purpose of our experiment is to unfold the direct link between levels of external socio-cognitive behavioral demand and inner-affective mechanisms. Through an ensemble learning task we attempted to exploit different behaviors of the selected base learners to enhance the accuracy of our overall learning system. Our aim is to show that defusers' physiological patterns differ between periods of longer and shorter replies and that there exist a range of stress levels across defusers.

6.1. Methodology

Feature Selection: Due to the fact that some of the features are highly correlated, we reduce the set with correlation feature selection (CFS). CFS selects features that correlate with the class label but are not correlated with previously selected features.

Ensemble Learning: To maintain diversity of base learners (Dietterich, 2002) we use two heterogeneous learner schemes: Voting (Shipp and Kuncheva, 2002) and Meta-learning (Wolpert, 1992). Under the Voting scheme, we combine the individual base by applying the average combination rule to the outputs. Meta-learning employs several base learners to get class predictions, which are then used by a meta-learning algorithm during the training phase to predict when the base learners are incorrect. Additionally, we comment on the ensemble size of the learners. Both ensemble schemes were built by combining the following base learners: K-Nearest Neighbor with K=5 (KNN)[1], Naive Bayes (NB), Decision Tree (DT), Random Tree (RT), Support Vector Machines with RBF kernel (SVM-RBF), Multi-layer Perceptron (MLP) and Random Forest (RF)(Breiman, 2001)[2]. The experiments were performed using leave-one-instance-out cross validation, where instance denotes a turn-taking behavioral response. We applied this approach for every defuser separately, for both conditions IB and CT respectively, as we wanted to investigate the unique individual trends of each defuser participant with respect to their behavior, their physiology and their their experimental condition (Ice Breaker or Control).

6.2. Results

The individual base learners and the ensemble learning methods chosen for our study are shown in Table 2.

Our experimental results range from 43.75% to 88.89%, suggesting that physiological signals contain information relevant to the amount of behavioral verbal replies. Additionally, we notice a great difference in performance across defusers, underlying once again the individual traits of every defuser. Particularly, the selected physiological cues of defusers P100, P102, P202, P203 and P204 appear to be more closely associated to the type of behavioral reply instances (short/long) compared to the corresponding patterns of P101, P103, P104 and P201 defusers.

[1] K=5 was empirically found to give better performance considering the limitation of small number of instances for 2 defusers (P104 and P202 with 9 and 10 instances respectively).

[2] This learner as a base one is robust and works relatively well without excessive need of meta parameter tuning.

		Base Learners							Ensemble Learners	
Participants	Condition	KNN(5)	NB	DT	RT	SVM-RBF	MLP	RF	Voting	Meta-learning
P100	IB	68.97	**75.86**	62.07	**75.86**	72.41	65.52	**75.86**	72.41	72.41
P101	IB	43.75	56.25	50.00	50.00	43.75	50.00	50.00	43.75	**62.50**
P102	IB	**80.00**	73.33	73.33	73.33	40.00	66.67	73.33	73.33	73.33
P103	IB	34.38	46.88	53.13	43.75	59.38	53.13	43.75	59.38	**62.50**
P104	IB	66.67	44.44	**88.89**	66.67	55.56	66.67	66.67	55.56	55.56
P200	CT	77.78	66.67	**88.89**	**88.89**	**88.89**	77.78	**88.89**	**88.89**	83.33
P201	CT	43.75	43.75	56.25	18.75	31.25	50.00	25.00	31.25	**62.50**
P202	CT	70.00	80.00	**90.00**	60.00	60.00	80.00	80.00	60.00	70.0
P203	CT	**88.00**	**88.00**	**88.00**	72.00	**88.00**	84.00	72.00	**88.00**	**88.00**
P204	CT	**77.78**	66.67	**77.78**	**77.78**	**77.78**	66.67	66.67	**77.78**	66.67

Table 2: The individual base learners and the ensemble learning methods chosen for ten defusers, five from the IB and CT conditions respectively during the "bomb defusion" task. The best method(s) for every participant is highlighted.

Ensemble learning performance and ensemble size: We notice that ensembles for the two different combination schemes either outperform the best individual base learner (P101, P103, P201) or reach similar performance with that (P200, P203, P204). After experimenting with a size ranging from 5 to 10 base learners, we present only those, whose combination determined the best approach towards building a stable diverse ensemble that generalizes well on the external and inner cues of individual. Furthermore, we have experimented with an odd and even number of ensemble size. Experimentally, we found that using an odd number for the ensemble size provides a higher learning performance. That could be explained if we consider that, when an even number of base learners is used, there is a potential for a tie when half of the base learners vote for one class while the other half vote for the opposite class.

Most frequently selected physiological features: We elaborate on the features presented in Table 1 in terms of their importance with respect to the "bomb defusion" task. We observe that for 50% of the defusers the most selected physiological signal is HR. Based on this, we assume that HR is associated with arousal levels and is of high importance for the examined task for this work. Regarding the remaining selected physiological features, we notice that these features range across defusers. This finding enhances the original assumption of uniqueness of individual personal traits across participants. Finally, the former finding is also aligned with our experimental results and observations that suggest that the "teammate prior familiarity" parameter does not have an impact on our task.

7. Discussion

As discussed in Section 6.2., there is a wide variability across defusers with respect to the given task. This observation indicates that there might be mechanisms triggered in defusers with high learning accuracy, reflected their physiological signals, which are not present in defusers with low learning performance (i.e. P101, P103, P201). It is also noteworthy that for these three defusers, the selected physiological feature is HR. To further elaborate on this tendency, we go through the audiovisual recordings and the HR signals. We notice that there is a difference in the arousal levels (i.e., stress) with respect to the type of behavioral replies (short/long) and that arousal affectivity is present both in short and long turn-taking responses, depending on the defuser.

More specifically, we come across examples of defusers who took a long time to respond after having given a wrong answer once and were asked to try again to confirm the bomb defusion steps. Hence, it appears that the task was a sufficiently stressful stimulus for them. In these long turn-taking examples, it is also reasonable to assume that high cognitive activity or stressor events occurred. At the same time, high levels of arousal are noticed in short turn-taking examples, in which for example the defuser uses words such ok/yes/no. This tendency is not aligned with the "bomb defusion" task, considering that we were expecting that short turn-taking examples would reflect low levels of arousal. On the contrary, our observation suggests that even though there may be no obvious (audible/visible) signals of arousal, physiological signals may provide a complementary, not overlaid though, view of a person's state. This finding is of particular importance, especially in cognitively demanding tasks in which one of the teammates manipulates the discussion and is also aligned with previous research studies (Tomaka et al., 1997; Gellatly and Meyer, 1992; Calkins and Fox, 2002).

8. Conclusions and Future Work

This study provides an analysis of physiological signals in a dyadic team "bomb defusion" scenario in association with their expressive behavioral cues. The results suggest that physiological responses convey information about the defuser's inner state. They also reflect the amount of the defuser's verbal responses with respect to a stimuli and can be further linked with the amount of underlying socio-cognitive activity, which is not always obvious through traditional observational methods. Last, we proposed two existing ensemble learning methods which are new to the field of generalizing external and inner cues of speakers, showing that these methods can yield improvements over traditional analysis methods.

One of the limitations of our study is its reliance on a small part of the corpus. Future plans include the analysis and discussion of the identified trends over the whole corpus, as well as the examination of the uniqueness of personal traits of every participant after switching roles (each team member was given the opportunity to be both the defuser and the instructor). Also, considering that this study relied on observational cues concerning turn taking dura-

tion measures, we believe that the examination of expressive cues with a more detailed analysis of participants' lexical features will provide an insight into whether these can be linked with their inner physiological signals. In terms of the lexical features, we would like to focus on the number of words, the length of the utterances, the number of laughs, the richness of the vocabulary as well as the use of backchannels in terms of short feedback such as "mmhmm", "yeah'.

Additionally, the investigation of singular pronouns (I, me, mine), assents (OK, yes), non-fluencies (hm, umm), fillers (I mean, you know) prepositions or words indicating prior familiarity could extend the pool of the used features. Apart from that, we intend to apply more advanced lexical modeling such as topic modeling, to better capture word usage, word choice and to unfold all aspects of the defuser's specific grammar employed in such stressful interactions. But, mostly we do believe that such an investigation could provide an insight with respect to the relevant vocabulary that is used in such particular tasks and the speaking style of every team member while sessions progress. This corpus serves as a technical springboard for developing dialogue agents that not only capture turn-taking behavior in a stressful task, but also underlying physiological states during the dyadic task.

9. References

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Calkins, S. D. and Fox, N. A. (2002). Self-regulatory processes in early personality development: A multilevel approach to the study of childhood social withdrawal and aggression. *Development and Psychopathology*, 14(3):477–498.

Cieslak, M. (2017). Moving ensemble averaging program. `https://github.com/mattcieslak/MEAP`.

Dennison, M., Neubauer, C., Passaro, T., Harrison, A., Scherer, S., and Khooshabeh, P. (2016). Using cardiovascular features to classify state changes during cooperation in a simulated bomb diffusal task. In *Proceedings of the Physiologically Aware Virtual Agents Workshop*. IEEE.

Dietterich, T. G. (2002). Ensemble learning. *The handbook of brain theory and neural networks*, 2:110–125.

El-Sheikh, M., Cummings, E. M., and Goetsch, V. L. (1989). Coping with adults' angry behavior: Behavioral, physiological, and verbal responses in preschoolers. *Developmental Psychology*, 25(4):490.

Gellatly, I. R. and Meyer, J. P. (1992). The effects of goal difficulty on physiological arousal, cognition, and task performance. *Journal of Applied Psychology*, 77(5):694.

Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., et al. (2011). Multiple classifier systems for the classification of audio-visual emotional states. *Affective Computing and Intelligent Interaction*, pages 359–368.

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Jones, G. R. and George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of management review*, 23(3):531–546.

McEwen, B. S. (2007). Physiology and neurobiology of stress and adaptation: central role of the brain. *Physiological reviews*, 87(3):873–904.

Murphy, L. R. (1996). Stress management in work settings: a critical review of the health effects. *American Journal of Health Promotion*, 11(2):112–135.

Neubauer, C., Woolley, J., Khooshabeh, P., and Scherer, S. (2016). Getting to know you: a multimodal investigation of team behavior and resilience to stress. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 193–200. ACM.

Newlin, D. B. and Levenson, R. W. (1979). Pre-ejection period: Measuring beta-adrenergic influences upon the heart. *Psychophysiology*, 16(6):546–552.

Prati, L. M., Ceasar, D., Ferris, G. R., Ammeter, A. P., and Buckley, M. R. (2003). Emotional intelligence, leadership effectiveness, and team outcomes. *International Journal of Organizational Analysis*, 11(1):21.

Raux, A. and Eskenazi, M. (2009). A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 629–637. ACL.

Richter, M. and Gendolla, G. (2009). The heart contracts to reward: monetary incentives and pre-ejection period. *Psychophysiology*, 46(3):451.

Schels, M. and Schwenker, F. (2010). A multiple classifier system approach for facial expressions in image sequences utilizing gmm supervectors. In *20th International Conference on Pattern Recognition*, pages 4251–4254. IEEE.

Schels, M., Glodek, M., Meudt, S., Schmidt, M., Hrabal, D., Böck, R., Walter, S., and Schwenker, F. (2012). Multi-modal classifier-fusion for the classification of emotional states in woz scenarios. In *1st International Conference on Affective and Pleasurable Design*, pages 5337–5346.

Scherer, S., Schwenker, F., and Palm, G. (2008). Emotion recognition from speech using multi-classifier systems and rbf-ensembles. *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*, pages 49–70.

Schuller, B., Müller, R., Lang, M., and Rigoll, G. (2005a). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Ninth European Conference on Speech Communication and Technology*.

Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., and Rigoll, G. (2005b). Speaker independent speech emotion recognition by ensemble classification. In *IEEE International Conference on Multimedia and Expo,(ICME) 2005.*, pages 864–867. IEEE.

Seery, M. D. (2011). Challenge or threat? cardiovascular indexes of resilience and vulnerability to potential

stress in humans. *Neuroscience & Biobehavioral Reviews*, 35(7):1603–1610.

Shipp, C. A. and Kuncheva, L. I. (2002). Relationships between combination methods and measures of diversity in combining classifiers. *Information fusion*, 3(2):135–148.

Spangler, D. P. and Friedman, B. H. (2015). Effortful control and resiliency exhibit different patterns of cardiac autonomic control. *International Journal of Psychophysiology*, 96(2):95–103.

Stein, F. (2001). Occupational stress, relaxation therapies, exercise and biofeedback. *Work: A Journal of Prevention, Assessment and Rehabilitation*, 17(3):235–245.

Tomaka, J., Blascovich, J., Kibler, J., and Ernst, J. M. (1997). Cognitive and physiological antecedents of threat and challenge appraisal. *Journal of Personality and Social Psychology*, 73(1):63.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.