TOWARDS PREDICTING PHYSIOLOGY FROM SPEECH DURING STRESSFUL CONVERSATIONS: HEART RATE AND RESPIRATORY SINUS ARRHYTHMIA

Arindam Jati[1], *Paula G. Williams*[2], *Brian Baucom*[2], *Panayiotis Georgiou*[1]

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

ABSTRACT

Being affected by mental stress during conversations might have a direct or indirect effect on our speech acoustics as well as on our physiological responses. This paper presents a study on finding the relationship between these two modalities, speech acoustics and physiology, during stressful conversations between humans. Heart rate and respiratory sinus arrhythmia have been considered as physiological variables in the present study. Two datasets, one from stress induction sessions and the other one from in-lab discussions of relationship conflicts between couples, have been analyzed. A series of experiments have been performed separately on the two datasets, as well as on the combined dataset. The research finds acoustic features that are significantly correlated with the physiological variables during stressful conversations. It also predicts the physiological signals from speech features through a nonlinear regression analysis. The results take us one step forward towards building an extremely non-intrusive and relatively inexpensive method of predicting physiological responses from speech, and thus detecting the presence and quantifying the intensity of stress during stressful conversations.

***Index Terms*—** Stress, speech, acoustics, physiology, heart rate, respiratory sinus arrhythmia

1. INTRODUCTION

Excessive stress can lead to a variety of physiological, psychological, and psychosomatic health conditions such as anxiety and depression [1], lack of proper physiological functioning causing degradation in performance, and even cardiovascular and cerebral diseases [2]. Automatic detection of stress can be very useful as it can warn the user and even help reduce stress-related health problems in the long run. It can also be helpful to psychologists during observational therapy as a diagnostic tool.

Effects of mental stress on physiology: Stress has a lot of short- and long-term effects on our physiology even though it has a psychological origin. Taelman *et. al.* [3] found significant change in Heart Rate (HR) and Heart Rate Variability (HRV) due to mental stress, which indicated strong potential for measuring stress levels from these signals. In [4], the authors studied the effect of stress originating from overcommitment and effort-reward imbalance at work on HR, blood pressure and vagal tone [4]. They discovered increased HR reactivity, increased systolic blood pressure and decreased vagal tone during stressful work. The relationship between stress and the respiratory pattern has also been studied quite thoroughly in the past [5, 6]. Sues *et. al.* [6] reported the occurrence of hyperventilation or over-breathing due to stress. Grossman [5] described the effect of respiration associated with stress responses on cardiovascular dysfunction. Porges [7] showed how Respiratory Sinus Arrhythmia (RSA), which is defined as the periodical alteration of heart rate in association with the phase of respiration [5], can be utilized as a technique to asses stress and vulnerability to stress.

Stress detection from physiology: The effect of mental stress on physiology has inspired a lot of researchers to build automated systems to detect stress by measuring physiological signals. For example, Healey and Picard [8] described methods to detect stress from direct measurement and analysis of different physiological signals like electrocardiogram, Electrodermal Activity (EDA) and respiration. Researchers of [9] utilized EDA for discriminating between stress and cognitive load in office environment. Choi and Gutierrez-Osuna [10] estimated the level of activation of sympathetic and parasympathetic nervous systems by analyzing HRV to detect the presence of stressful events.

Stress detection from speech: Although these methods provide quite satisfactory accuracy for stress detection, they need intrusive and in some cases invasive methods to acquire the physiological signals. This inspired the speech scientists to harness the prosodic and emotional cues of speech to predict stress [12]. A hefty amount of work has been done in this field using the SUSAS dataset [13]. The nonlinear Teager Energy Operator (TEO) [13] on acoustic features was found to be useful for detecting stress. A classification experiment was performed in [14] for detecting stress from speech for drivers engaged with cognitive load. [15] also did stress classification experiments from both speech and galvanic skin response. A Gaussian Mixture Model (GMM) based framework was proposed in [16] to detect stress in speech under physical load. Lu *et. al.* [17] built a system for detecting stress in varying acoustics environments using smartphones.

Relationship between speech and physiology: As we saw the previous examples, the speech and physiology modalities have been well studied and employed to detect mental stress. Another interesting area of research is to explore the relationship between these two modalities. Predicting physiological responses from speech acoustics during stressful conversations can give some insight on how a psychological variable (the reason of stress) can lead to both physiological and vocal activations, and how these two are related.

Finding acoustic features that are related to physiological signals can also help in the development of a multi-modal stress detection systems. Moreover, estimating the raw values of physiological signals can provide higher resolution quantitative metrics for the intensity of stress as opposed to just detection of the presence of stress.

However, to the best of our knowledge, only a few attempts have been made in the past to tackle the problem of predicting physiological signals only from audio. [18] tried to predict HR from pronunciation of vowels. Schuller and his colleagues [19] analyzed correlation, regression and classification results for the tasks of vowel pronunciation and reading out a sentence loud with and without physical load. A recent study [20] tried to classify change in the direction of HR from acoustic features for conversation with an artificial dialog system. Their newer study [21] proposed a regression analysis to

Table 1. Datasets

Name	#Sessions	#Males	#Females	Total duration
SI	54	29	25	6.20 hours
CI	226	115[†]	111[†]	22.30 hours

predict HR values from audio using the same corpus.

Present study: The present study investigates the relationship between acoustic features and physiological signals (HR and RSA) in terms of correlation and regression analyses for spontaneous stressful conversations between humans in two separate datasets. The main novelties and importance of this paper are the followings: *i)* To the best of our knowledge, this is the first ever attempt to explore the relationship between the two aforementioned modalities for conversations between humans; *ii)* and the first study to predict RSA from speech, and identify the acoustic features that facilitate that prediction; *iii)* The proposed work employs two very distinct datasets and thus addresses issues related to robustness of the acoustic features across different domains; *iv)* and our work provides regression analysis on the actual value of the physiological signals thus better addresses the strengths and limitations of the physiological-speech connection during unconstrained, fluent and real conversations. This could in turn elucidate the possibility of building an audio-based automated real-time stress or physiology monitoring system [19].

2. DATASETS

Two different datasets, Stress Induction (SI) and Couples' Interaction (CI), have been utilized for the present analysis. The summary of the two datasets are reported in Table 1. Note that the 'total duration' field tabulates total audio duration of all sessions before performing any preprocessing.

2.1. Stress Induction (SI) dataset

The original dataset [22] was collected from 98 young adults who *re-experienced* their top two *stressors* (or stressful events in life) in a semi-structured Social Competence Interview [22]. HR, RSA, blood pressure and some other physiological measures had been collected along with speech. Baseline [22] or resting physiological responses of all the participants had also been collected. We use only the re-experience part of the interview (around 4-6 minutes per stressor) for this study. We denote one session as a combination of the two re-experience parts from two stressors. So, the average duration of each session is around 8-12 minutes. We work on a subset of the original dataset. Our dataset has 54 sessions. All the sessions have manual transcriptions of speech, that we utilize to extract participant's speech from a dyadic conversation between participant and interviewer as described in Section 3.1. For every session, average values of HR and RSA over the entire session have been used in the present analysis. More details about the study and data can be found in [22].

2.2. Couples' Interaction (CI) dataset

Sixty married, community couples were recruited in Salt Lake City, UT for participation in a study of communication and emotion in marriage. The recruitment inclusion criteria included: the couples being legally married for at least one year and living together, both spouses speaking fluent English, and being between the ages of 18 and 60. Recruited couples were married for a mean of 10.0 years (SD = 7.60) at the beginning of the study. The mean age of the recruited wives was 41.6 years (SD = 8.59), and the mean age of the husbands was 43.5 years (SD = 8.74). The mean number of years of education was 17.0 for both the wives and husbands (SD = 3.23 for wives, SD = 3.17 for husbands). The majority of the participants were Caucasian (wives: 76.1%, husbands: 79.1%); other well-represented ethnicities included African American (wives: 8.2%, husbands: 6.7%), Asian or Pacific Islander (wives: 4.5%, husbands: 6.0%), and Latina/Latino (wives: 5.2%, husbands: 5.2%).

Each couple received up to 26 sessions of therapy over the course of one year. As part of the study, research staff had couples select two current, serious relationship problems, one chosen by each partner, and then had them engage in two dyadic discussions in which they were instructed to try to understand and resolve these respective relationship problems. There was no therapist or research staff present during these sessions, and the couple interacted for ten minutes about the wife's chosen topic and ten minutes about the husband's chosen topic; these two ten-minute sessions were considered separate and analyzed separately. The problem-solving interactions were recorded at three points in time across the study: pre-therapy, the 26-week assessment, and the two-year post-therapy assessment. The sessions from the pre-therapy are employed in this paper.

The audio-video data consist of a split-screen video (29.97 fps) and a single channel of far-field audio recorded from the video camera microphone (16 kHz, 16-bit). Since the data were originally only intended for manual coding by experts, the recording conditions were not ideal for automatic analysis; the video angles, microphone placement, and background noise varied across couples and across sessions. We separate the audio streams of two individuals from the dyadic conversation of a couple by speaker diarization, as will be discussed in Section 3.1, and then analyze each of them with the average physiological responses (HR and RSA) for the corresponding person. Therefore, two sessions for each of the 120 different people (60 couples) contribute to a total of 240 audio streams. But we discard very short audio streams that mainly occurred whenever a particular person spoke very little (and the speaker diarization algorithm failed to separate that speaker) in a conversation. This results in total 226 sessions for further analysis.

For both the datasets, we have multiple baseline physiological values for every participant. We use the average of all baseline values for normalization purpose (will be described in Section 4.2).

3. METHODOLOGY

3.1. Audio preprocessing

For the CI dataset, each session has been passed through a denoising module (implemented using VOICEBOX[1]) which does speech enhancement using MMSE estimate of spectral amplitude [23]. The SI dataset didn't require any denoising because of relatively lower noise levels.

Two different techniques have been employed on the two datasets to reject the silence regions from audio streams. For the SI dataset, this is done by forced alignment of the audio with the transcript using the Gentle toolkit[2]. For the CI dataset, a robust LSTM based Voice Activity detection (VAD) implementation [24] from OpenSMILE [25] has been used since no transcripts were available for this dataset.

After that, speaker diarization has been performed to delineate the utterances of the two speakers in a given session. Forced alignment has been used on the SI dataset (since it has manual transcripts), and a hierarchical agglomerative clustering implementation

[†] Number of unique males/females = 60.

[1] `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

[2] `https://lowerquality.com/gentle/`

Table 2. Pearson's correlation (all correlations are statistically significant, i.e. $p < 0.05$) between the physiological variables (HR or RSA) and the best correlated feature (indicated inside parenthesis)

Gender	**SI dataset**		**CI dataset**	
	RSA	HR	RSA	HR
Male	-0.40 *(mean falling slope of loudness)*	-0.53 *(mean bandwidth of 3rd formant)*	-0.40 *(coefficient of variation of bandwidth of 1st formant)*	0.36 *(coefficient of variation of bandwidth of 1st formant)*
Female	-0.42 *(20th percentile of loudness)*	0.55 *(range of 20th to 80th percentile of loudness)*	0.40 *(voiced segments per second)*	0.35 *(mean harmonics to noise ratio)*

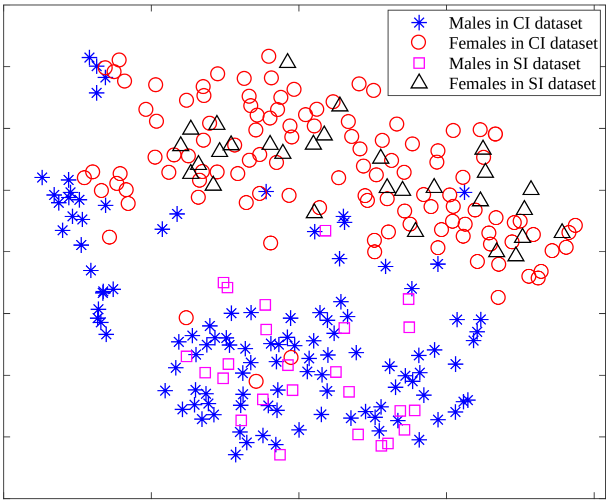

Fig. 1. t-SNE plot of the acoustic features on both datasets.

from LIUM speaker diarization tool[3] has been employed on the CI dataset (since it doesn't have manual transcripts).

After that, speaker identity has been assigned to each of the two audio streams. For the SI dataset, this is done by forced alignment because we have manual transcripts annotated with speaker identity. For the CI dataset, we automatically separated speech streams of husband and wife (they are from different genders for every couple in the dataset) based on their average pitch profile over the entire session [26].

3.2. Acoustic feature extraction

We have extracted 88 dimensional extended GeMAPS (or eGeMAPS) features [27] from speech over the whole session using OpenSMILE toolkit. For the SI dataset, the features have been extracted from the participant's speech in a given session. For the CI dataset, they have been extracted separately from husband's and wife's speech over the entire session. The eGeMAPS features mainly consist of statistical functionals (mean and coefficient of variation [27]) of frequency related parameters like pitch, jitter, and format frequencies; energy/amplitude related parameters like shimmer, loudness, and harmonics to noise ratio; spectral balance parameters like alpha ratio, Hammarberg index, and harmonic differences; some temporal features like rate of loudness, and mean length of voiced regions; and finally some cepstral features like MFCC and spectral flux. The full list of features is as described in [27].

Table 3. RMSE for regressing raw physiological variables (HR or RSA)

Gender	**SI dataset**		**CI dataset**	
	RSA	HR	RSA	HR
Male	1.81	12.61	1.14	8.99
Female	2.01	11.28	1.22	11.94

3.3. Gender dependent models

During experimentation we have observed that computing correlation separately for male and female speakers gives much better results for both the datasets. We have noticed similar behavior for the regression analysis as well. To find the reason behind this, we computed the t-SNE transformation [28] of all the acoustic features in both datasets to visualize them in a much lower dimension. Figure 1 shows the t-SNE result as a 2-dimensional scatter plot. We can see clear clusters were formed within males and females from both datasets, possibly because of fundamental differences between some of the acoustics features (for example pitch) among men and women. This inspired us to use gender dependent models for further analysis.

3.4. Regression of physiological variables

We have employed AdaBoost regressor [29] with decision tree regressor as base estimator for estimating raw or normalized values of RSA and HR from acoustic features. All the 88 features have been used by the regression model. 5-fold stratified cross validation has been done. During regression we have made sure not to include speech of the same speaker in both training and testing data. As we have seen in Section 2, the SI dataset has one session per speaker (so simple 5-fold cross validation works for this dataset), while the CI dataset might have at most two sessions per speaker. For the CI dataset, we have employed 5-fold cross validation such that no two folds have speech from the same speaker or even same couple. Optimal number of base estimators and the learning rate [29] of the regressor have been chosen through 3-fold cross validation on the training set and searching the parameter space by grid search. The whole process has been repeated 5 times to get a better estimate of test error.

4. EXPERIMENTS AND RESULTS

4.1. Analysis of two datasets separately

Pearson's correlation coefficients have been calculated between the acoustic features and the average values of the physiological signals (separately for HR and RSA) over all sessions. Table 2 lists the best correlations obtained along with the name of the corresponding features for both the datasets. All the p-values for the correlations are statistically significant (two-tailed t-test). Maximum (in absolute sense) correlations for RSA (-0.42) and HR (0.55) have been observed for females in SI dataset. We should also notice that different features are dominating across the two datasets for the same physiological variable in the same gender group. This might be be-

[3] `http://www-lium.univ-lemans.fr/diarization/doku.php/Welcome`

Table 4. Pearson's correlation (all correlations are statistically significant, i.e. $p < 0.05$) between raw or normalized physiological variables (HR or RSA) and the best correlated feature (indicated inside parenthesis) for combined dataset

Gender	**Raw**		**Normalized**	
	RSA	HR	RSA	HR
Male	0.22 *(mean bandwidth of 3rd formant)*	-0.24 *(coefficient of variation of shimmer)*	0.33 *(coefficient of variation of bandwidth of 3rd formant)*	0.43 *(mean alpha ratio)*
Female	-0.22 *(mean length of voiced segments in seconds)*	-0.20 *(mean of shimmer)*	0.19 *(standard deviation of falling slope of pitch)*	0.35 *(50th percentile of loudness)*

Table 5. RMSE for regressing raw or normalized physiological variables (HR or RSA) in combined dataset

Gender	**Raw**		**Normalized**	
	RSA	HR	RSA	HR
Male	1.13	9.90	1.07	8.82
Female	1.42	11.82	0.89	7.99

cause of the difference in the tasks (re-experiencing stressors in SI dataset versus discussing about relationship conflicts in CI dataset) the speakers are performing in the two datasets. For example, during re-experiencing a stressor in front of an interviewer, one participant might not be very emotionally or vocally aroused. On the other hand, the same speaker might show strong arousal while discussing relationship conflicts under stress with his/her spouse.

Table 3 shows the regression results in terms of Root Mean Squared Errors (RMSE). It might be useful to know that the range of raw values (minimum, maximum) for RSA and HR over both the datasets are (2.04, 8.89) and (50.0, 112.08) respectively. Although we got overall better correlation for SI dataset (Table 2), the regression results on SI dataset are a little bit worse than that of CI dataset. We suspect the reason to be the far fewer number of training samples in SI dataset. It is worth mentioning here that the observed correlations and RMSE values are similar (or sometimes better for the case of HR) to previous study [19] (please see the speaker independent (LOSO) case in that paper) although the speaking tasks are very different.

4.2. Combining two datasets

To see the robustness of the results we also analyzed the correlation and regression performance after combining both the datasets. The 'Raw' column of Table 4 presents Perason's correlation values between the best correlated acoustic feature and the raw physiological values. We can see the correlations drop by a large margin from the values we obtained separately in two datasets. One major reason for this might be coming from different distributions of the physiological signals in the two datasets because of the inherent difference between the tasks the users are doing there (as we discussed in Section 4.1). So, to tackle the situation, we normalize the physiological signals for every user by subtracting the corresponding average baseline physiological values (as discussed in Section 2). Now we get a boost in the correlations (please see 'Normalized' column of Table 4) except for RSA of females. The best correlations for RSA (0.33) and HR (0.43) have been observed for males in the combined datasets.

Table 5 shows the RMSE values for regressing the raw and normalized physiological values on the combined dataset. We can see better performance for predicting raw values (even though correlations degraded) than what we obtained only on SI dataset. This has possibly happened because of having more training samples. Note that we can not directly compare the 'Normalized' RMSE with the 'Raw' RMSE because of the range difference of their values due to normalization.

5. CONCLUSION

In this study, we tried to find the relationship between acoustics and physiology during stressful conversations between humans. Employing gender dependent models was found to be more useful than gender-independent ones. Separate experiments on two datasets helped us identify acoustic features that are significantly correlated with the physiological responses. We noticed a degradation in correlation when we combined two datasets, but per-speaker normalization of the physiological variables enhanced the performance. While regressing the physiological variables separately on two datasets, we observed overall (3 out of 4 cases) better performance for the dataset with more participants. Possibly due to the same reason, we achieved satisfactory RMSE values for the combined dataset.

This study helps us infer that it is possible to find some acoustic features that are significantly correlated with the physiological signals under the conditions of both our datasets (Section 2). This seems to agree with the hypotheses from Section 1 that the stress has an effect on both modalities, speech and physiology. The regression results give a better picture about how accurately we can predict physiological signals (RSA and HR for this study) from speech acoustics for stressful conversations between humans. This also sheds light on our idea of quantifying the intensity of stress instead of just detecting it.

In the future, we are planning to investigate on finding better acoustic features. We will also apply deep learning models and exploit any available temporal pattern in the speech signal that could help us predicting physiological responses more accurately. Further, the connection between physiology and acoustics can be studied without any human labeling, so it opens up possibilities for larger data collections and more robustly associating the two via advanced modeling.

6. ACKNOWLEDGMENT

The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Research Program under Award No. W81XWH-15-1-0632. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.

7. REFERENCES

[1] Sheldon Cohen, Ronald C Kessler, and Lynn Underwood Gordon, *Measuring stress: A guide for health and social scientists*, Oxford University Press on Demand, 1997.

[2] John T Cacioppo, Louis G Tassinary, and Gary G Berntson, "Psychophysiological science," *Handbook of psychophysiology*, vol. 2, pp. 3–23, 2000.

[3] Joachim Taelman, Steven Vandeput, Arthur Spaepen, and Sabine Van Huffel, "Influence of mental stress on heart rate

and heart rate variability," in *4th European conference of the international federation for medical and biological engineering*. Springer, 2009, pp. 1366–1369.

[4] Tanja GM Vrijkotte, Lorenz JP Van Doornen, and Eco JC De Geus, "Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability," *Hypertension*, vol. 35, no. 4, pp. 880–886, 2000.

[5] Paul Grossman, "Respiration, stress, and cardiovascular function," *Psychophysiology*, vol. 20, no. 3, pp. 284–300, 1983.

[6] William M Suess, A Barney Alexander, Deborah D Smith, Helga W Sweeney, and Richard J Marion, "The effects of psychological stress on respiration: a preliminary study of anxiety and hyperventilation," *Psychophysiology*, vol. 17, no. 6, pp. 535–540, 1980.

[7] Stephen W Porges, "Cardiac vagal tone: a physiological index of stress," *Neuroscience & Biobehavioral Reviews*, vol. 19, no. 2, pp. 225–233, 1995.

[8] Jennifer A Healey and Rosalind W Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[9] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *IEEE Transactions on information technology in biomedicine*, vol. 14, no. 2, pp. 410–417, 2010.

[10] Jongyoon Choi and Ricardo Gutierrez-Osuna, "Using heart rate monitors to detect mental stress," in *Wearable and Implantable Body Sensor Networks, 2009. BSN 2009. Sixth International Workshop on*. IEEE, 2009, pp. 219–223.

[11] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss, "Activity-aware mental stress detection using physiological sensors," in *International Conference on Mobile Computing, Applications, and Services*. Springer, 2010, pp. 282–301.

[12] Margaret Lech and Ling He, "Stress and emotion recognition using acoustic speech analysis," in *Mental Health Informatics*, pp. 163–184. Springer, 2014.

[13] Guojun Zhou, John HL Hansen, and James F Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 9, no. 3, pp. 201–216, 2001.

[14] Raul Fernandez and Rosalind W Picard, "Modeling drivers speech under stress," *Speech communication*, vol. 40, no. 1, pp. 145–159, 2003.

[15] Hindra Kurniawan, Alexandr V Maslov, and Mykola Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. IEEE, 2013, pp. 209–214.

[16] Sanjay A Patil and John HL Hansen, "Detection of speech under physical stress: Model development, sensor selection, and feature fusion," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[17] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 351–360.

[18] D Skopin and S Baglikov, "Heartbeat feature extraction from vowel speech signal using 2d spectrum representation," in *Proc. of the 4th International Conference on Information Technology (ICIT), Amman, Jordan*, 2009, p. 6.

[19] Björn Schuller, Felix Friedmann, and Florian Eyben, "Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7219–7223.

[20] Andreas Tsiartas, Andreas Kathol, Elizabeth Shriberg, Massimiliano de Zambotti, and Adrian Willoughby, "Prediction of heart rate changes from speech features during interaction with a misbehaving dialog system," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] Jennifer Smith, Andreas Tsiartas, Elizabeth Shriberg, Andreas Kathol, Adrian Willoughby, and Massimiliano de Zambotti, "Analysis and prediction of heart rate using speech features from natural speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 989–993.

[22] Paula G Williams, Matthew R Cribbet, Holly K Rau, Heather E Gunn, and Laura A Czajkowski, "The effects of poor sleep on cognitive, affective, and physiological responses to a laboratory stressor," *Annals of Behavioral Medicine*, vol. 46, no. 1, pp. 40–51, 2013.

[23] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[24] Florian Eyben, Felix Weninger, Stefano Squartini, and Björn Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 483–487.

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

[26] Eluned S Parris and Michael J Carey, "Language independent gender identification," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. IEEE, 1996, vol. 2, pp. 685–688.

[27] 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., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[28] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[29] Harris Drucker, "Improving regressors using boosting techniques," in *ICML*, 1997, vol. 97, pp. 107–115.