VOCE Corpus: Ecologically Collected Speech Annotated with Physiological and Psychological Stress Assessments

Ana Aguiar[†], Mariana Kaiseler[§], Mariana Cunha[†], Jorge Silva[†], Hugo Meinedo[‡], Pedro R. Almeida[¶]
[†] Instituto de Telecomunicações, Porto, Portugal
[§] Carnegie Faculty, Leeds Metropolitan University, Leeds, UK
[‡] INESC-ID, Lisbon, Portugal
[¶] University of Porto - Faculty of Law, Porto, Portugal
[†] ana.aguiar@fe.up.pt, [§] m.h.kaiseler@leedsmet.ac.uk [‡] hugo.meinedo@l2f.inesc-id.pt, [¶] palmeida@direito.up.pt

Abstract

Public speaking is a widely requested professional skill, and at the same time an activity that causes one of the most common adult phobias (Miller and Stone, 2009). It is also known that the study of stress under laboratory conditions, as it is most commonly done, may provide only limited ecological validity (Wilhelm and Grossman, 2010). Previously, we introduced an inter-disciplinary methodology to enable collecting a large amount of recordings under consistent conditions (Aguiar et al., 2013). This paper introduces the VOCE corpus of speech annotated with stress indicators under naturalistic public speaking (PS) settings. The novelty of this corpus is that the recordings are carried out in objectively stressful PS situations, as recommended in (Zanstra and Johnston, 2011). The current database contains a total of 38 recordings, 13 of which contain full psychologic and physiologic annotation. We show that the collected recordings validate the assumptions of the methodology, namely that participants experience stress during the PS events. We describe the various metrics that can be used for physiologic and psychologic annotation, and we characterise the sample collected so far, providing evidence that demographics do not affect the relevant psychologic or physiologic annotation. The collection activities are on-going, and we expect to increase the number of complete recordings in the corpus to 30 by June 2014.

Keywords: Stress, psychophysiologic, speech, corpus

1. Introduction

Public speaking (PS) is an important component across professional settings, and it has been suggested that the fear of PS, called glossophobia, is the most common adult phobia (Miller and Stone, 2009). One of the main challenges in PS is the negative experience of stress while speaking, and its detrimental effects on speech performance.

The study of stress under laboratory conditions may provide only limited ecological validity (Wilhelm and Grossman, 2010). Hence, we aim at collecting a large corpus of speech annotated with stress in objectively stressful situations, as recommended in (Zanstra and Johnston, 2011). Such methods are known as ecological momentary assessment, or momentary studies (Stone and Shiffman, 2002), or also as naturalistic settings in the Engineering community. With this purpose, we previously presented an inter-disciplinary methodology that enables the collection of a large amount of recordings under consistent conditions (Aguiar et al., 2013). The current work describes the full methodology and preliminary results of a larger research project, VOCE[1], that ultimately aims to classify stress from speech alone for the purpose of designing computer assisted voice coach applications.

The contributions of this paper are to 1) briefly describe the data collection methodology and platform; 2) characterise the current VOCE corpus, and make it available to the community; 3) provide supporting evidence for the assumption that the naturalistic settings chosen are perceived as stressful; 4) provide guidance on which physiologic metrics are better suited for stress annotation of data gathered under naturalistic conditions. For this purpose, we collected 38 PS recordings following the designed methodology, which build up the first release of the VOCE corpus. Of these, 13 include the full psychologic and physiologic annotation. The data collection is currently ongoing, and we expect to increase the number of complete recordings in the corpus to 30 by June 2014. A final release of the corpus is planned for March 2015.

2. Related Work

Emotion recognition from speech has been previously addressed, e.g. (Yang et al., 2012; Ververidis and Kotropoulos, 2006; Scherer, 2003), however little is known about the study of stress and PS events. In an attempt to further investigate this are, (Lu et al., 2012) tested a mobile phone platform for stress detection from speech is designed and developed for a wide range of indoor and outdoor environments. However, the authors did not assess whether the participants are actually experiencing stress. Instead they assume that stress was experienced when a particular event occurs without contemplating the individual appraisal, as suggested by Lazarus (Lazarus and Folkman, 1984). Our methodology addresses this limitation through the use of validated questionnaires and physiological indexes of stress. Another important limitation is that neither the tools nor the corpus were made available. In (Zuo et al., 2012), the authors described the methodology for collection of a corpus of multilingual speech annotated with stress, where the participants were university students. In their study, stress was induced through questions. We follow a naturalistic approach in which participants feel stressed due to the real world situation. Moreover, our corpus innovates by assessing individual ratings of stress based on

[1] paginas.fe.up.pt/~voce

self-reports and physiological measures. Self-assessment measures are used to validate the subjective experience of stress, while physiologic data provides fine-grained stress annotation during the speech.

Another aspect of related work refers to speech databases and corpora. Computer-based detection of human emotions has been studied using three types of emotional materials (Hansen et al., 2000; Scherer, 2003; Ververidis and Kotropoulos, 2006): acted, spontaneous and elicited emotions. Acted emotions are less ambiguous, as actors are trained to express emotions accurately. Spontaneous emotions are more difficult to gather and to initially classify because humans frequently do not experience a single emotion, but a mixture of emotions with different intensities. Elicited emotions are induced and can be as difficult to classify as spontaneous emotions (Kessous et al., 2010; Vogt et al., 2008). Several databases for automatic speech-based emotion recognition have recently been set up, with different number and diversity of subjects, as well as variety of tagged emotions (a thorough study can be found in (Ververidis and Kotropoulos, 2006)). Although one of those databases focuses on stress, SUSAS (Hansen et al., 2000), it is based on single word utterances, and 96% come from aircraft communication, restricting generalization of the results to the PS setting. More recently, Zuo at al. presented a growing corpus of stress annotated speech (Zuo et al., 2012), but their corpus differs from ours in the points mentioned above.

3. Ecologic Methodology

Participants in the current study are student volunteers and the main part of the recordings takes place in an actual PS event that is part of the academic curriculum (e.g., presentations of coursework, research seminars). All participants are native European portuguese speakers. The participants complete informed consent forms and health questionnaires at the moment of volunteering.

Psychological stress was assessed using the portuguese version (Ponciano et al., 2005) of the State Trait Anxiety Inventory (STAI) (Spielberger et al., 1983). The instrument has successfully been used to evaluate stress/no stress conditions, e.g. (Kaiseler et al., 2012). It consists of 20 questions (state anxiety), in which participants are required to rate their feelings by answering "How are you feeling right now?" questions using a 4-point Likert scale anchored at 1 = "Not at all" and 4 = "Very much". Good psychometric properties (reliability and fit indicators) were reported for the this scale (Spielberger et al., 1983). Additionally, demographic and health questionnaires were administered to participants as a means to trace eventual bias caused by, e.g. legal drugs, or physical or mental illness.

We use Heart Rate Variability (HRV) as a physiological index of stress. The analysis of the HRV provides well-known and accepted estimators of parasympathetic activity and have been thoroughly used as correlates of psychological stress (Berntson and Cacioppo, 2004; Allen et al., 1991). Both time and frequency domain measures of the HRV exist, whereby time domain measures are more adequate for long term analysis and frequency domain measures more adequate to short-term analysis of stress (of The European Society of Cardiology et al., 1996). Common time domain measures are the average and standard deviation of the heart rate (HR) and of the intervals between R peaks of consecutive QRS complexes of the electrocardiogram (RR intervals). Frequency domain measures are the power in pre-defined bands of the power spectral density of the sequence of RR intervals, namely the low frequency (LF, 0.04 to 0.15 Hz) and the high frequency (HF, 0.15 to 0.4 Hz) bands, and relationships between them, namely the autonomic balance, measured as LF/HF and the normalised LF, LF/(LF+HF). We refer the reader to (of The European Society of Cardiology et al., 1996) for more detailed information on the HRV measures, their meaning and calculation and to (Berntson and Cacioppo, 2004; Allen et al., 1991) for more information on the relationship between stress and HRV.

Our methodology consists in using the psychologic self-assessment to validate that the PS event is experienced as stressful, and in collecting speech synchronised with physiologic sensor data (RR intervals) during the PS event. Since the first is a momentary assessment immediately before the PS event, it does not provide stress/no-stress differentiation at utterance level. Hence, we use the physiologic sensor data to obtain utterance level granularity stress annotation during the speech. Another aspect that must be considered is that HRV measures are adequate only for intra-subject comparison, and cannot be used for comparison between subjects. For this reason, a Baseline recording of each subject if necessary. To accommodate this, each recording consists of 3 sub-recordings:

Baseline recorded at least 24 h before the PS event, consists of demographic questionnaire and STAI, reading a standard text and heart monitor sensor data;

Experiment recorded no more than 30 min before the PS event, consists of STAI, reading the same standard text as before and heart monitor sensor data;

Event recorded during the actual PS event, consists of free speaking and simultaneous heart monitoring.

We refer the reader to (Aguiar et al., 2013) for more details on the methodology and recording procedure.

4. Platform

We developed a dedicated, easy-to-use platform that implements and enforces the same work-flow across all recordings (sequence of actions, questionnaire order, speech volume adjustment, verification of sensor readings, etc.), thus standardizing the procedure and reducing variations due to varying recording conditions. The platform uses only commercial off-the-shelf (COTS) hardware: a wireless headset microphone (AKG PW45 SPORT SET), an A/D converter working at 44 KHz sampling rate and 24 bits/sample (M-AUDIO FAST TRACK MKII), a Zephyr HxM BT[2] heart rate monitor, an Android smartphone and a laptop. The participants self-assessed whether the recording paraphernalia

[2] http://www.zephyr-technology.com/products/hxm-bluetooth-heart-rate-monitor/

impacted stress experience, and answers indicated that it was not relevant.

The platform was designed to be easily re-used elsewhere in other projects. Besides COTS hardware, we use Java language and questionnaires are stored in XML files read at runtime for the sake of adaptability to other questions or languages.Please contact the authors for more information on how to obtain it.

The synchronization of the physiological sensors with speech signal is guaranteed by a heart-beat message sent every 5 s from the laptop to the smartphone, which the latter uses to timestamp the values received from the heart sensor. Start and stop of sensor recording is also synchronized with speech start and stop by messages sent by the laptop. Thus, physiological sensor data is recorded with the same time reference as the voice, so HRV values can be matched to utterances. Collected data for each recording is initially stored in the file system of the laptop and uploaded on-demand to a storage platform at the end of a recording.

Finally, we have created a web platform for crowd-sourcing assessments by others, not involved in the collection processes or PS events. The platform asks the user to listen to the Baseline and Experiment voice samples from a random participant and asks which one sounds calmer. This platform is available at `http://176.111.105.16/webplatform/index.php`, and the annotations are available as part of the VOCE corpus.

5. VOCE Corpus

The current version of the VOCE corpus (first release) consists of a collection of 38 raw recordings as described above, adding up to 78 min of Baseline, 73.6 min of Experiment and 487 min of Event free speech, with accompanying metadata (demographic and health questionnaires). Speakers are 38 students from the University of Porto, aged 19 to 49.

Only 22 of those recordings are complete, although 28 have all data for Baseline and Experiment. This was caused by some inconsistencies in the initial version of the platform, which have now been corrected. Nevertheless, the incomplete recordings are valuable. For example, the STAI scores, the demographic and health information are available for all recordings. Also, 9 of the 16 incomplete recordings have complete Baseline and Experiment sub-recordings, which are speech recordings of the same text at two moments in time where the stress levels are different. These incomplete recordings are useful, for example, for validation of the methodology assumptions.

Further, of the 22 complete recordings, only 13 have the RR intervals as physiologic data, while the other 9 have only the heart rate data. For the first, it is possible to calculate all the HRV measures mentioned in Section 3., while for the latter only the average and standard deviation of the HR can be obtained.

Finally, we have listened to all full recordings, and chosen the ones with best audio quality. The audio quality impacts strongly the performance of automatic segmentation algorithms and other speech recognition software that VOCE corpus users may wish to apply to the recordings. Due to the naturalistic collection environments, audio quality varies due to factors like the varying acoustics of the room, background noise, etc. Hence, we also identify those as a separate group.

Table 1 summarises the 4 groups of recordings.

Group	Description	Nr of individuals
A	all speakers observed	38
B	speakers with HR measures	22
C	speakers with RR measures	13
D	speakers with best audios	20

Table 1: Sample groups.

We publish the raw recordings, instead of utterances and physiologic features, to provide independence from the post-processing algorithms and tools that we chose to use in our further work, namely the speech segmentation and the physiologic sensor processing. As such, the VOCE corpus is generic and not polluted by eventual inaccuracies of those algorithms or implementations.

The corpus consists of the metadata, raw audio (.wav) and sensor files (XML) for each recording. The metadata available for each recording consists of the following fields:

recID recording random identifier;

age speaker age;

gender speaker gender;

eventDescription type of event;

scientificArea speaker study field;

health self-assessment of health condition;

physical regular physical activity (yes/no);

physicalActivity type of regular physical activity;

physicalTimes regularity of physical activity (days/week);

disease speaker has a heart disease (yes/no);

diseaseDescription which disease;

drugs speaker regularly takes legal drugs (yes/no);

drugsDescription which legal drugs;

tobacco speaker smokes (yes/no);

tobaccoDay how many cigarettes per day;

coffee speaker regularly drinks coffee (yes/no);

coffeeDay how many coffees per day;

StaiScoreQ1 Baseline STAI score, calculated from answers at the beginning of the Baseline sub-recording;

StaiScoreQ2 Experiment STAI score, calculated from answers at the beginning of the Experiment sub-recording;

State recording status, which is either ”complete” or described the available components;

Warnings warnings regarding audio quality, added after listening to each individual recording.

The sensor data for each sub-recording is stored in an xml file with the following scheme:

```
<Baseline Start="1369645794">
  <Description>
    <Key Name="ecg">77, 1369645795</Key>
    <Key Name="ts">20113, 1369645795</Key>
    <Key Name="ts">19361, 1369645795</Key>
    <Key Name="ecg">78, 1369645797</Key>
    <Key Name="ts">20849, 1369645797</Key>
    ...
     </Description>
</Baseline>
```

with the following meanings:

(sub-recording) Start UTC timestamp of the recording begin;

Key Name="ecg" is the value of the heart rate, this is a value averaged over some window and after some filtering, for which details are hidden in the Zephyr heart rate monitor;

Key Name="ts" are the timestamps of the R peaks in the QRS complexes detected by the Zephyr device since the last stored value[3], measured according to an internal clock of 16 bits (these values are unfiltered).

We also provide a set of files with the sequence of RR intervals extracted from the sensor files.
The corpus can be downloaded from the links available at `http://paginas.fe.up.pt/~voce/articles.html`. A readme.txt describes which recordings belong to which group in Table 1.

6. Stress Experience Validation

This section presents results that validate psychological and physiological stress during the PS event through comparison of Baseline and Experiment. Additionally, we evaluate the possible associations between the variations and the demographic parameters. Analysis is separated in psychologic self-assessment and the physiologic assessment as shown in sections 6.1.. and 6.2.., respectively .

6.1. Psychological stress

For validating the psychological stress during the event, we use the perceived stress self-assessment on the Baseline and Experiment. The self-assessment is obtained from processing the results of the STAI into a score that varies between 20 and 80, whereby higher scores indicate greater anxiety (Spielberger et al., 1983).
To test this parameter behaviour we applied the Mann - Whitney -Wilcoxon signed rank test that has three types of hypothesis tests:

$$H_0 : F_{Bas} = F_{Exp} // H_1 : F_{Bas} \neq F_{Exp} \quad (1)$$

$$H_0 : F_{Bas} \leq F_{Exp} // H_1 : F_{Bas} > F_{Exp} \quad (2)$$

$$H_0 : F_{Bas} \geq F_{Exp} // H_1 : F_{Bas} < F_{Exp}, \quad (3)$$

[3]Unfortunately, due to an implementation error, these values are not stored sequentially, and must be ordered before processing.

Group	STAI values		$H_0 : F_{Bas} \leq F_{Exp}$	
	Baseline	Experiment	z-value	p.value
A	Mean=31.9 SD=7.8	Mean=37.6 SD=9.8	3.37	0.0004
B	Mean=30.6 SD=7.7	Mean=35.1 SD=8.5	2.52	0.006
C	Mean= 31.5 SD=7.4	Mean=33.6 SD=7.9	1.1	0.136
D	Mean=30.7 SD=8.7	Mean=39.7 SD=10.3	2.60	0.004

Table 2: Variation of STAI and significance level of hypothesis tests.

where $F_{Bas} < F_{Exp} \Rightarrow P(S_B < x) < P(S_E < x) \Rightarrow$ means higher values in Baseline than Experiment.
We tested all types of tests in all groups and the most significant results were obtained for type 2) in all cases, i.e., when the alternative is *Higher values in Experiment than in Baseline*. The results of these tests are summarised in Table 2.
Group A participants show significantly higher scores ($p < 0.01$) in state anxiety during the Experiment condition compared with Baseline. So, we can conclude that participants experience more anxiety before the Experiment when compared to the Baseline.
For group B the results were also significant but less divergent. Participants statistical values of STAI in Baseline compared with Experiment diverge 5 points, and there is a significant ($p < 0.05$) trend of higher values in Experiment compared with Baseline.
For group C, the difference between STAI means is only 2.1 and the MWW test of type 2) is only indicative ($p < 0.15$). Based on the significant results obtained for all other groups, we strongly believe that this is due to the reduced power in the analysis caused by the small sample size.
For group D, the difference between means of STAI is higher than all previous groups and the MWW test of type 2) is highly significant ($p < 0.01$). This reveals that the speakers with recordings of higher quality are experiencing higher levels of anxiety during the Experiment (prior to the PS event) compared with the Baseline conditions.

6.2. Physiologic stress

We will separate the analysis of physiological stress measures into temporal and spectral measures. The temporal HR measures can be calculated for the participants in group B plus 6 others for which we have the HR data for Baseline and Experiment, whereas the Avgnn, SDnn, and spectral measured can only be calculated for the participants in group C. The summary of the statistics for all HRV measures can be found in Table 3 and the results of the Wilcoxon-Signed-Rank tests applied to the variations of the HRV measures for the 3 hypothesis mentioned above are shown in Table 4.

6.3. Temporal HRV Measures

The analysis of the physiologic data collected for group B shows changes in temporal HRV measures: the aver-

Measure	Baseline		Experiment		Monotony
	Mean	StDev	Mean	StDev	
HRavg *	93.59	18.84	99.35	19.96	increases
HRstd *	5.7	7.2	4.17	2.75	decreases
Avnn **	0.7135	0.1524	0.6938	0.1468	decreases
SDnn **	0.06342	0.0268	0.0529	0.0216	decreases
LF **	0.0020856	0.00232	0.001293	-0.00100	decreases
HF **	0.001242	0.001727	0.000906	0.0010118	decreases
LF/HF **	2.623959	2.107055	3.249573	4.106916	increases
LF/(LF+HF) **	0.63097	0.2273172	0.6416758	0.1654956	increases
* - data from group B+6; ** - data from group C					

Table 3: Basic statistics of HRV temporal and spectral measures

Parameters	Alternative		
	$H_1 : F_{Bas} \neq F_{Exp}$	$H_1 : F_{Bas} > F_{Exp}$	$H_1 : F_{Bas} < F_{Exp}$
STAI *	0.0118	0.0059	0.9945
avgHR *	**0.1041**	**0.0520**	0.9506
stdHR *	0.2850	0.8628	**0.1425**
Avnn **	0.6247	0.7119	0.3123
SDnn **	0.4016	0.8181	0.2008
LF **	0.3636	0.8360	0.1818
HF **	0.8339	0.4169	0.6101
LF/HF **	0.6749	0.6876	0.3375
$LF/(LF+HF)$**	0.5761	0.7353	0.2880
* - data from group B+6; ** - data from group C			

Table 4: p-values of MWW test for different STAI and HR measures. Recall that $F_{Bas} < F_{Exp} \Rightarrow P(S_B < x) < P(S_E < x) \Rightarrow$ higher values in Baseline than Experiment.

age HR increases significantly (z-value=1.63, $p = 0.052$) and the HR variance (avg=5.7 and std=7.2 in Baseline vs avg=4.17 and std=2.75 in Experiment) decreases (z-value=-1.07, $p < 0.15$) indicatively. Avnn and SDnn also decrease from Baseline to Experiment, but not significantly (z-value=0.489, p=0.312), probably due to low power caused by the small sample size. These results are in accordance with the expected physiologic response to social stressors.

6.4. Spectral HRV Measures

The sequences of RR intervals were processed using the Physionet[4] toolkit, a validated and freely available tool for extracting HRV metrics from RR interval sequences (Goldberger et al., 2000). Specifically, we used the *get_hrv* function with standard parameters and outlier filters. For the results presented in this section, we fed the RR interval sequence for the whole partial recording (Baseline or Experiment) into the tool, whereby they varied between lengths of 58 and 169 for Baseline, and 60 and 170 for Experiment. The tool provides LF and HF powers, and LF/HF and LF/(LF+HF) for each Baseline and Experiment subrecordings. The latter are both suggested measures for assessing autonomic balance (of The European Society of Cardiology et al., 1996). The variations of the LF and HF powers were also tested, since they may prove to be more reliable measure sin the out-of-the-lab settings that we use. Specifically, since the HF measure may be noisy due to physical activity, causing noise also in the composed measures, the LF power may be a better marker of periods of higher autonomic imbalance. The Wilcoxon-Signed-Rank test on the spectral HRV measures are performed only for group C.

The average of LF decreases from Baseline to Experiment, having significant variability in Baseline and a shorter relative deviance in Experiment. Although the MWW test is not significant (z-value=0.9085, p=0.18), it has the lowest p-value and is the most likely to become significative with increasing sample size. HF average has significant variation in both situations and of the same order of magnitude, therefore no significant result comes from MWW test (z-value$<$0.21, $p >$0.4). LF/HF has deviances equal or higher than means in Baseline and in Experiment, making it impossible to conclude about its divergence. It is correlated with age ($\rho = -0.649$) and health ($\rho = 0.520$). Finally, LF/(LF+HF) increases from Baseline to Experiment but not significantly (z-value=0.5591, p=0.2880).

From these results, we conclude that the LF spectral HRV measure is the most likely to show significant variation between Baseline and Experiment as the sample grows. Hence, it is probably the most adequate of the covered HRV measures to use as indicator of physiologic stress for speech annotation in our case.

7. Conclusions

In this paper, we describe a methodology for collecting speech annotated with psychological and physiological measures of stress among college students in naturalistic

[4] www.physionet.org

settings of public speaking. The psychologic stress annotations are used to validate that the participants experience high levels of anxiety during the public speaking events. . The physiologic measures are used to automatically annotate the free speech. The recordings collected so far validate the assumptions of the methodology at least indicatively, both in terms of psychologic self-assessment and physiologic activation. We expect that both the psychological and physiological stress measure variations will become significant as the sample size increases.

This publication makes available to the research community a first release of the VOCE corpus with 38 raw recordings and a total of 638 min of speech annotated with physiologic sensors from which stress measures can be obtained. We are confident that this is a useful contribution to the research community aiming to further understand the effects of psychophysiological stress levels on speech. The platform developed to implement the proposed methodology is based on commercial-off-the-shelf devices and java language, and can be easily used by other research groups in other projects anywhere. We encourage the reader to try it out by contacting the authors.

Future work focuses on going the corpus, classification of stress from speech features, feature selection for real-time implementations of stress detection from speech. We also aim at developing adequate biofeedback mechanisms that will provide a speaker with useful real-time feedback to improve his public speaking skills.

Due to the interdisciplinary nature of the project, main aims and naturalistic approach used, it was not possible to contemplate a more qualitative analysis of the stress appraisal for the different participants (Lazarus and Folkman, 1984). Hence, we recommend that future research in this area should combine quantitative and qualitative measures of stress.

8. Acknowledgements

The authors thank the VOCE project advisory board for insightful discussions and comments: Anibal Ferreira, Jaime Cardoso, Isabel Trancoso, Cristina Queirós, Miguel Coimbra. Further, we thank the students that have collaborated and contributed to the data gathering: Tiago Borba, Mariana Pereira, Paula Fortuna.

This work was supported by Fundação para a Ciência e a Tecnologia, through projects VOCE (PTDC/EEA-ELC/121018/2010) and PEst-OE/EEI/LA0008/2013.

9. References

Aguiar, A., Kaiseler, M., Meinedo, H., Abrudan, T., and Almeida, P. R. (2013). Speech Stress Assessment using Physiological and Psychological Measures. In *Proceedings of the 2nd ACM Workshop on Mobile Systems for Computational Social Science at Ubicomp*.

Allen, M., Boquet, A., and Shelley, K. (1991). Cluster analyses of cardiovascular responsivity to three laboratory stressors. *Psychosomatic Medicine*, 53:272–288.

Berntson, G. and Cacioppo, J., (2004). *Dynamic Electrocardiography*, chapter Heart rate variability: Stress and psychiatric conditions, pages 57–64. Future.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C.-K., and Stanley, H. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101.

Hansen, J., Swail, C., South, A., Moore, R., Steeneken, H., Cupples, E., Anderson, T., Vloeberghs, C., Trancoso, I., and Verlinde, P. (2000). The impact of speech under 'stress' on military speech technology. Technical Report RTO-TR-10 AC/323(IST)TP/5 IST/TG-01, NATO Research and Technology Organization.

Kaiseler, M., Polman, R., and Nicholls, A. (2012). Think aloud: Gender differences in appraisal and coping with stress during the execution of a complex motor task. *International Journal of Sport and Exercise Psychology*, 10(4):1–15.

Kessous, L., Castellano, G., , and Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3(1).

Lazarus, R. S. and Folkman, S. (1984). *Stress, appraisal and coping*. Springer.

Lu, H., Rabbi, M., Chittaranjan, G. T., Frauendorfer, D., Mast, M. S., Campbell, A. T., Gatica-Perez, D., and Choudhury, T. (2012). StressSense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones. In *Proceedings of ACM Ubicomp*.

Miller, T. and Stone, D. (2009). Public speaking apprehension (psa),motivation, and affect among accounting majors: A proof-of-concept intervention. *Issues in Accounting Education*.

of The European Society of Cardiology, T., of Pacing, T. N. A. S., and Electrophysiology. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17:354–381.

Ponciano, E., Loureiro, L., Pereira, A., and Spielberger, C. (2005). Características psicométricas e estrutura factorial do tai de spielberger em estudantes universitários. In Motta, A. P. . E., editor, *Actas do Congresso Nacional Acção Social e Aconselhamento Psicológico no Ensino Superior: Investigação e Intervenção*, pages 315–322.

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., and Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Consulting Psychologists Press.

Stone, A. and Shiffman, S. (2002). Capturing momentary, self- report data: A proposal for reporting guidelines. *Annals of Behavior Medicine*, 24(3):236–243.

Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

Vogt, T., Andre, E., and Wagner, J., (2008). *Affect and Human Emotions in HCI*, volume 4868 of *LCNS*, chapter Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realization. Springer.

Wilhelm, F. and Grossman, P. (2010). Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biological Psychology*, 84:552–569.

Yang, N., Muraleedharan, R., Kohl, J., Demirkol, I., Heinzelman, W., and Sturge-Apple, M. (2012). Speech-based emotion classification using multiclass svm with hybrid kernel and thresholding fusion. In *Proceedings of the 4th IEEE Workshop on Spoken Language Technology*, Dec.

Zanstra, Y. and Johnston, D. (2011). Cardiovascular reactivity in real life settings: Measurement, mechanisms and meaning. *Biological Psychology*, 86:98–105.

Zuo, X., Li, T., and Fung, P. (2012). A Multilingual Natural Stress Emotion Database. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.