Speech Stress Assessment using Physiological and Psychological Measures

A. Aguiar
Instituto de Telecomunicações,
Porto, Portugal
ana.aguiar@fe.up.pt

P. R. Almeida
University of Porto - Faculty of Law, School of Criminology
Porto, Portugal
palmeida@direito.up.pt

M. Kaiseler
University of Porto - Faculty of Psychology and Educational Sciences
Porto, Portugal
mkaiseler@fpce.up.pt

H. Meinedo
INESC-ID
Lisboa, Portugal
hugo.meinedo@l2f.inesc-id.pt

T. E. Abrudan
Instituto de Telecomunicações,
Porto, Portugal
tabrudan@fe.up.pt

UbiComp '13 Adjunct, Sept 8-12, 2013, Zurich, Switzerland.
ACM 978-1-4503-2139-6/13/09...$15.00.

Abstract

Emotional stress is commonly experienced while speaking in public, producing changes to the various speech productions subsystems, affecting the speech signal in predictable ways and being easily conveyed to listeners. Speech stress indicators, however, are typically studied under laboratory settings, allowing little generalization to real life settings. To bridge this gap, we propose an interdisciplinary approach to assess speech stress during public speaking events, based on a platform that records speech simultaneously annotated with physiological and psychological measures. This approach enables the collection of a large corpus of annotated speech in ecological settings, i.e. in objectively stressing situations. We also propose and implement a methodology to assess listeners evaluation of stress including psychologists, and overall public.

The platform has been in use for the past 5 months, and we have collected over 30 complete samples after the initial iterative development procedure. Preliminary results indicate that the proposed user-friendly platform is an accurate and robust method to collect annotated speech under ecological settings that can be processed to obtain speech stress indicators. The findings will be used primarily in the design of computer and mobile assisted voice coaching applications, but the outreach extends to

mobile emotion sensing for individuals and crowds.

Author Keywords

Stress Assessment, Speech, Physiological sensors, Methodology, Platform

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

General Terms

Stress detection, Platform Optional section to be included in your final version.

Introduction

Public speaking (PS) is an important component across professional settings. In agreement with this idea [10] suggested that the clinically diagnosable fear of PS, called glossophobia, is the most common adult phobia (irrational fear). One of the main challenges in PS is the negative experience of stress while speaking, and the detrimental effects on speech performance. Due to the numerous situations and contexts where it occurs, the experience of stress in speech is an imperative issue for investigation and a portal for intervention programmers. Despite this need, previous research in this area has mainly been conducted under laboratory conditions. This fact can be probably associated with the methodological challenges of assessing stress in PS under ecological conditions. Although results from laboratory studies develop the knowledge in the area, their provide limited generalization to real life based assessments of stress [18]. To overcome this limitation it has been recommended that stress responses should be studied in relation to objectively stressful situations [20]. Particularly, the 21st century science, recommends research methods such as ecological momentary assessment, also known as momentary studies [16] to assess stress. Following this recommendation, in this work, we develop tools for assessing psychological and physiological responses of stress collected during real world live speech.

Speech samples are collected in public speaking events, and annotated using non-intrusive, and user friendly physiological sensor data. The collection work-flow includes the stress evaluation collected by self-reports and physiologic measures. It has been suggested by Mehrabian that how we say something is more important than what we say to the listener's (e.g. psychologists, general public) perception of credibility and sincerity [9]. To further test this argument, we will test others' perception of speech stress. Hence, this work will include physiological annotation of stress using various bio-medical sensors and self-report measures to assess psychological stress and listeners evaluation of stress including psychologists, and overall public. We plan to make the annotated corpus of speech samples that we are building with the methodology described in this paper available to the scientific community as soon as a significant size has been achieved. Following data collection and stress categorization, stress related speech feature extraction and selection will follow, and classification will be conducted. Our ultimate goal is to classify stress from speech alone for the purpose of designing computer assisted voice coaching applications. Although emotion recognition from speech has been addressed before [19, 17, 13], we focus specifically on stress and PS events.

The contributions of this paper are: 1) the design of an ecological methodology for assessing stress in speech; 2) a technological platform to enable large-scale collection of speech data annotated with psychological assessments

and physiological sensors on affordable hardware; 3) preliminary results obtained with the proposed platform that validate the proposed methodology.

Ecologic Methodology

Designing a methodology to record speech annotated with other indicators of stress in ecological setting targeting a voice coaching application raised a set of challenges:

1. verifying that participants actually experience psychological stress during the chosen ecological settings;
2. recording indicators of stress with fine time granularity that enable a fine-grained annotation of speech, so that feedback may be provided by a computer-assisted voice coaching application;
3. synchronizing the stress indicators with speech, so that the fine-grained annotation of speech in ecological settings with physiological stress will serve as the ground truth to train classifiers that can detect stress from voice features alone;
4. collecting a large amount of samples to train a classifier with good generalization properties;
5. collecting samples in a least possible obtrusive manner, to minimize the bias caused by the methodology.

To address the first challenge, we use State-Trait Anxiety Inventory (STAI) to measure psychological stress at the time of the PS event, and compared the results with a baseline measurements collected for each participant more than 24h away from the event. The second challenge requires a more intricate approach, because STAI questionnaires assess psychological stress on a coarse time frame, i.e. they assess stress at the time of the event, but do not enable fine-grained assessment of stress variations during the actual PS event, as would be needed for a coaching application. So, we resource to physiological measures of stress, as these enable a fine-grained evaluation of the stress experienced by the speaker. The third challenge is addressed by recording the voice and physiological sensors with the same time reference, for which purpose we designed a dedicated platform.The 4th and 5th challenges are softer, in the sense that they are less technical, but they played a critical role in the design of the dedicated platform. Collection of a large amount of samples requires a recording setup that is easy to follow, to be usable by a wide range of "gatherers". At the same time, it should guarantee that the same procedure is followed in each recording, to avoid variability caused by the sample collection process. Finally, the 5th requirement impacts the design of the procedure to be followed and the hardware choices.

In the rest of this section, we describe the psychological and physiological measures and finish with the procedure designed. The technical details of the platform are presented at the end of this section.

Psychological Measures

Psychological stress was assessed using the portuguese version [12] of the STAI [15]. The instrument has successfully been used to evaluate stress/no stress conditions, e.g. [8]. It consists of 20 questions, in which participants were asked to answer "How are you feeling right now?". The scale uses a 4-point Likert type response scale anchored at 1 = Not at all and 4 = Very much. Because the present study focuses on state anxiety only the first 20 items were considered. Good

psychometric properties (reliability and fit indicators) have been reported for the this scale [15]. Additionally, demographic and health questionnaires were provided to participants as means to trace eventual bias caused by, e.g. legal drugs, or physical or mental illness.

Physiological Measures
For assessing physiological stress we opted for the heart rate variability (HRV), which is a well-known and accepted measure of the activation of the parasympathetic system and of experience of stress [3, 1]. There are time and frequency domain measures of the HRV, whereby time domain measures are more adequate for long term analysis and frequency domain measures more adequate to short-term analysis of stress. Average and standard deviation of the heart rate are common time domain measures. The most commonly used frequency domain measure is the rate of the power in the low frequency to the high frequency bands of the spectrum of the sequence of intervals between consecutive RR peaks, which are the intervals between R peaks of consecutive QRS complexes of the electrocardiogram. We refer the reader to [11] for more detailed information on the measures and to [3, 1] for more information on the relationship between stress and HRV.

Recording Procedure
Participants in the study fulfill informed consent forms and health questionnaires at volunteering time, and recording sessions are scheduled. On the day prior (24h) to the PS event, the gatherer meets the participant and Baseline test is conducted. This includes the fulfillment of a demographic questionnaire and the STAI in the platform. Following this procedure, the participant puts on the microphone and attaches the heart monitor belt to the chest, and the researcher connects the equipment with the platform. Participants are instructed to do a voice test, to verify the volume of the recorded sound. Following this stage, participants are instructed to read a standard text, consisting of 3 paragraphs of emotions free content and taking approximately 90 s, while the heart is being monitored. The Baseline test takes approximately 8 min to complete. In the day of the PS event, the gatherer mets the participant again 60 to 30 min before the event. The equipment is set up as in the Baseline, and the same procedure as in the Baseline tests is followed, including reading of the same standardized text while monitoring the heart. This stage takes approximately 8 min to complete. Following this stage, the participant fulfills a new STAI in the platform. Finally, the Event recording phase starts, consisting of the voice recording of the full presentation, with simultaneous heart monitoring.

Summarizing, a recording consists of 3 sub-recordings: Baseline, at least 24 h from the PS event, consisting of demographic questionnaire and STAI, standard text reading and heart monitoring; Experiment, no more than 30 min from the PS event, consisting of STAI, standard text reading and heart monitoring; and the Event recording, consisting of free speaking and simultaneous heart monitoring. The STAI in Baseline and Experiment are used to validate that the participant is aroused in the PS event with respect to his normal state. The heart monitor data in Baseline and Experiment is used to validate that HRV is a physiological indicator of psychological stress on a timescale of 90 s. The voice recording in Baseline and Experiment is used to search for features that are the best indicators of psychological stress. Finally, the physiological sensors in the Event will be used to tag instants of higher arousal in the participant's speech, serving as the ground truth in the training of a classifier based on the chosen features that

will be the base of a computer aided coaching application.

Platform

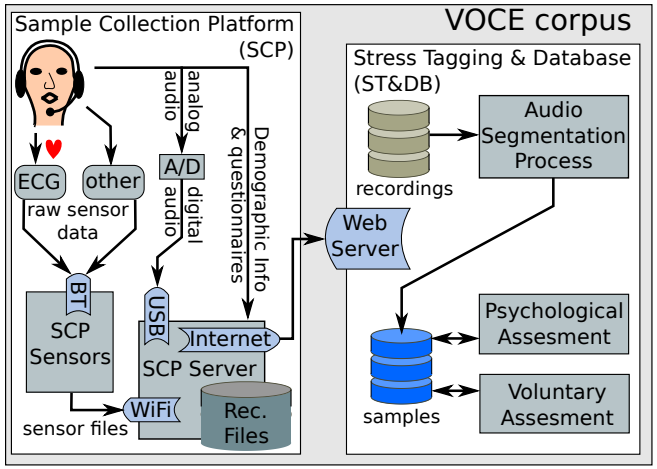

Figure 1: Dedicated platform for large scale collection of PS event recordings according to methodology designed.

We developed a dedicated, easy-to-use platform that implements the full work-flow from data recording to their final storage in a server, using commercial off-the-shelf hardware. To guarantee user-friendliness, the platform was developed iteratively, in development cycles interspersed with test recordings carried out by non tech savvy gatherers. The proposed architecture is illustrated in Figure Figure 1, and consists of two main modules: the Sample Collection Platform (SCP), and the Stress Tagging and Database (STDB).

Sample Collection Platform
The SCP consists of a distributed application built on a client-server architecture on a smartphone and a notebook, both portable devices. SCP Sensors and the SCP Server are the two applications, the former running on an Android OS smartphone is responsible for the collection of physiological sensor data over Bluetooth , while the latter, developed in Java and platform independent, controls the workflow, records the voice and collects the files gathered in the smartphone. The smartphone and netbook are connected by a WiFi link, on a hotspot created by the Android smartphone, so that the platform can be used anywhere independently of external connectivity. The choice of decentralizing physiological sensor collection to a smartphone was made to account for the limited range of Bluetooth communication and the fact that in PS events the gatherer may not be able to be located close to the speaker. The SCP Server application implements the procedure described above in a strict manner, guaranteeing the consistency of the recordings independently of the gatherer.

Collected data for each recording is initially stored in the file system of the netbook where SCP Server runs, and includes: answers to demographic questionnaire (1) and psychological questionnaires stored in XML files (Baseline and Experiment, totaling 2), audio WAV files (Baseline, Experiment and Event, totaling 3) and heart sensor data stored in XML files (idem). The XML files used comply with the ExMARALDA[1] standard for voice corpora. The SCP Sensor application also implements a client to the STDB service, enabling the on-demand upload of complete recordings to the storage platform.

Voice Sensor
To record the speech, we chose an AKG PW45 SPORT SET, which is a wireless headset microphone with pocket transmitter working on the UHF B1 frequency range. Although this choice pays a small cost of intrusiveness,

[1] http://www.exmaralda.org/en_index.htm

the headset microphone provides very low variability in the distance between the mouth and the microphone during the speech, minimizing signal power variations due to microphone position. The speech is then passed to an M-AUDIO FAST TRACK MKII A/D converter, working at 44 KHz sampling rate and 24 bits/sample.

SCP Server displays screens that allow the gatherer to adjust the gain controls and microphone position so that the recorded voice levels are within a range that will not cause clipping of the signal during the event. The speaker is presented with a set of sentences that contain sounds likely to cause such clippings and is requested to read the sentence until the recorded signal keeps inside the reference levels, which were empirically determined to be 45% of the maximum quantization level. In this phase, we ask the gatherers and speakers to change only the gain in the A/D converter, as it is the latest in the chain, keeping the previous gains to a minimum, so as to minimize the noise.

Physiological Sensor Annotation
Since we aim at collecting a large amount recordings in ecological settings, we had to choose a physiological sensor that is as non-intrusive as possible, to minimize the impact that the data collection setup may have on the event itself. Hence, we opted for the Zephyr HxM BT[2], a heart sensor as commonly used for sports and other lifestyle tracking, consisting of a chest strap and a small device that plugs to it can transmit data wirelessly over Bluetooth. The device has an open API that can be easily integrated in Android applications, and returns heart rate and the timestamps of the most recent 15 R peaks with 1 ms resolution, at a frequency of 1 Hz.

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

The sequence of RR intervals can be obtained from the R peak timestamps, and the physiological stress indicator HRV calculated. Since the R peak timestamps are recorded in the same time reference as the voice, each HRV value for each 90 s window can be matched to a speech segment.

Synchronization
The synchronization of the physiological sensors with the voice is guaranteed by a heart-beat message sent every 5 s from the SCP Server to the SCP Sensors application that the latter uses to timestamp the values received from the heart sensor. Moreover, the start and stop of sensor recording is synchronized with start and stop of the speech by messages sent by the SCP Server. Upon reception of a stop message, SCP Sensors sends the XML file with the sensor data to the SCP Server.

Stress Tagging and Database
The Stress Tagging and Database module is also part of the corpus acquisition, concentrating the storage and post-processing of the raw data recorded. Post-processing has two goals: validating that the PS event causes stress, and annotating the speech with the physiological sensors.

For validating the psychological stress during the event, we use two types of perceived stress assessment on the Baseline and Experiment: self-assessment and other's assessment. The self-assessment is obtained from processing the results of the STAI into a STAI score that varies between 20 and 80, whereby higher scores indicate greater anxiety [15]. The other's assessment will be crowd-sourced using a web platform that presents the user, e.g psychologist and overall public, with the Baseline and Experiment voice samples from one participant and asks which one sounds calmer. This platform is available at http://176.111.105.16/webplatform/index.php. Since

data collection procedures started recently, not enough assessments have been provided yet, limiting conclusions in the current paper. Nevertheless, we welcome the reader to try it out.

The audio files are segmented into utterances according to the full stop punctuation marks automatically recovered using the system described in [2]. These speech utterances represent the individual corpus elements, for which speech features are calculated that can be input into a classifier in a posterior phase. Hence, for each utterance we need a physiological stress assessment that will serve as a ground truth and can be obtained from the physiological annotations of the recordings.

The physiologic stress annotation is obtained from processing the heart rate (HR) recorded with the speech in windows of 100 s. We obtain the average and standard deviation of HR as time domain measures. Additionally, we obtain the RR interval in each second, calculate the spectrum using FFT, and calculate the HRV by dividing the power in the LF band (0.04–0.15 Hz) by the power in the HF band (0.15–0.4 Hz) [11]. In this fashion, we obtain an HRV value for each 100 s window, corresponding to one value for each Baseline and Experiment, and one value for all utterances that overlap with the time window for which the HRV value is calculated.

The database of the VOCE corpus will consist of one table characterizing the recording, one table containing the psychological assessments (STAI scores and others' scores) for Baseline and Experiment, three tables containing the start and duration of the utterances (Baseline, Experiment and Event) and the physiological stress assessments of each utterance.

Preliminary Results

Extensive testing of the proposed platform was conducted in real scenarios and its functionality verified. Until now, we have successfully collected 13 full recordings, but the results presented here for speech include only data from 10 of those. These results tendencially validate our methodology but do not all achieve statistical significance due to the reduced number of recordings, which we continue gathering. The participants in this first phase were asked to self-assess the impact of the recording paraphernalia in their stress level, and the answers indicate that it was not relevant. In the next sections, we present preliminary results that validate psychological and physiological stress through comparison of Baseline and Experiment data, as well as the feasibility of detecting stress from speech alone.

Psychological and Physiological Stress

STAI scores were analysed using the statistical Package for the Social Sciences (SPSS, version 19). A Wilcoxon signed-rank test showed that participants present significantly higher scores in state anxiety during the Experiment (Mean=43.62; StDev= 10.03) compared with Baseline condition (Mean=35.54; StDev = 8.32), achieving a z-value of 2.044 that corresponds to a confidence level higher that 5%. As expected, this fact is likely to be associated with the unpleasant state experienced by the participants prior to the PS event.

Analysis of the physiological data collected so far showed a significant increase in average HR and a tendency of increase in the autonomic balance, measured as the ratio of LF to HF powers, during Experiment when compared to Baseline, indicating a higher sympathetic activation and a decrease of the parasympathetic activity, consistent with the experience of stress. Specifically, we obtained

z-values of 2.696 and 1.148 from the Wilcoxon signed rank test, corresponding to confidence levels higher than 1% and 15%, respectively.

Speech Stress Recognition

We performed a preliminary experiment of stress determination from speech using the available Baseline and Experiment data recordings. For each of the 161 automatically determined utterances a feature vector containing 6125 speech features was extracted using the Opensmile toolkit [4]. This toolkit is capable of extracting a very wide range of speech features and has been applied with success in a number of paralinguistic classification tasks [14]. The 6125 features extracted were obtained by applying segment-level statistics (means, moments, distances) over a set of energy, spectral and voicing related frame-level features. From this large vector of utterance-level features a subset of 35 was chosen by performing automatic feature selection using the WEKA toolkit [5]. The selection algorithm [6] evaluates the worth of a subset of features by considering their individual predictive ability along with the degree of redundancy between them. Using the reduced subset of features and choosing the Experiment utterances as positive examples (containing stress) and the Baseline utterances as negative examples a binary classifier was trained in WEKA with the Random Subspaces algorithm [7] and 10-fold cross validation. This method consists of multiple decision trees constructed systematically by randomly selecting subsets of components from the feature vector. The final trained model has 10 decision trees each with 6 pseudorandomly selected speech features. The classification accuracy obtained when evaluated using all the utterances is 76.4%.

Conclusions

We propose an interdisciplinary methodology to assess stress in speech in ecological settings, namely in public speaking events, combining physiological and psychological annotations with speech recordings. We developed a technological platform that implements the methodology, enabling the collection of a large amount of recordings, which will be post-processed into a corpus of stress annotated utterances. Preliminary analysis of psychological results verify that participants experienced significantly higher levels of anxiety in the Experiment phase compared with the Baseline phase. Physiological sensors also indicate the presence of stress, although we do not achieve significant results yet. Hence, our findings show that the proposed method was successful in collecting speech stress in public speaking under real world conditions. A preliminary attempt at binary detection of stress in speech features shows that it is possible to detect stress in speech from a reduced set of features with an accuracy of 76%. The collected data will be made available to the scientific community. Future work will focus on increasing the volume of the corpus to achieve significant results on physiological as well as psychological sensors, as well as designing a reduced feature set to detect momentary stress in speech and applying the results to a mobile voice coaching application. Applications of the results extend further into the use of stress detection from speech in mobile applications for well being in other scenarios and social science research.

Acknowledgements

The authors would like to thank M. Coimbra, C. Queiroz, I. Trancoso, andJ.P. Cunha for insightful discussions.

This work was supported by Fundação para a Ciência e a Tecnologia, through projects VOCE

(PTDC/EEA-ELC/121018/2010) and
PEst-OE/EEI/LA0021/2011.

References

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

[2] Batista, F., Moniz, H., Trancoso, I., and Mamede, N. Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *Trans. Audio, Speech and Lang. Proc. 20*, 2 (Feb. 2012), 474–485.

[3] Berntson, G., and Cacioppo, J. *Dynamic Electrocardiography*. Future, 2004, ch. Heart rate variability: Stress and psychiatric conditions, 57–64.

[4] Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, MM '10, ACM (New York, NY, USA, 2010), 1459–1462.

[5] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The weka data mining software: an update. *SIGKDD Explor. Newsl. 11*, 1 (Nov. 2009), 10–18.

[6] Hall, M. A. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[7] Ho, T. K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell. 20*, 8 (Aug. 1998), 832–844.

[8] Kaiseler, M., Polman, R., and Nicholls, A. 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 (2012), 1–15.

[9] Mehrabian, A. *Silent messages: Implicit communication of emotions and attitudes*. Belmont: Wadsworth, 1981.

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

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

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

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

[14] Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., and Weiss, B. The INTERSPEECH 2012 Speaker Trait Challenge. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)* (2012).

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

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

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

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

[19] Yang, N., Muraleedharan, R., Kohl, J., Demirkol, I., Heinzelman, W., and Sturge-Apple, M. 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 2012).

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