Stress Detection from Speech and Galvanic Skin Response Signals

Hindra Kurniawan[1], Alexandr V. Maslov[1,2], Mykola Pechenizkiy[1]
[1]*Department of Computer Science, TU Eindhoven, the Netherlands*
hindra.kurniawan@gmail.com, m.pechenizkiy@tue.nl
[2]*Department of MIT, University of Jyväskylä, Finland*
alexandr.maslov@jyu.fi

Abstract

The problem of stress-management has been receiving an increasing attention in related research communities due to a wider recognition of potential problems caused by chronic stress and due to the recent developments of technologies providing non-intrusive ways of collecting continuously objective measurements to monitor person's stress level. Experimental studies have shown already that stress level can be judged based on the analysis of Galvanic Skin Response (GSR) and speech signals. In this paper we investigate how classification techniques can be used to automatically determine periods of acute stress relying on information contained in GSR and/or speech of a person.

1 Introduction

Chronic stress has become a serious problem affecting different life situations and carrying a wide range of health-related diseases, including cardiovascular disease, cerebrovascular disease, diabetes, and immune deficiencies [5]. In health-care and well-being research the problem of stress-management has been receiving an increasing attention [1, 2, 8]. Technologies that automatically recognize stress can become a powerful tool to motivate people to adjust their behavior and lifestyle to achieve a better stress balance. Technology such as sensors can be used for obtaining an objective measurement of stress level, which afterwards, machine learning technique is employed to build a model for stress detection and recognition.

Change detection [3] and classification [16] are two general types of approaches allowing to indicate whether person was stressed over a particular time period. In this paper we study how classification techniques can be used to automatically determine periods of acute stress in the offline settings. For this purpose we constructed a benchmark dataset by collecting GSR and voice recordings from ten people during a controlled study. Then we trained classifiers using feature extracted from GSR and/or speech signals. The results of our experiments show that for the constructed benchmark 1) SVM classifiers reach 92% stress detection generalization accuracy, 2) features extracted from speech have higher predictive power than GSR features but at the same time are more person dependent, i.e. they do not allow classifiers to generalize well to those people for whom no data were available during the model training phase, and 3) combining information from both signals together is not trivial, i.e. learning a classifier or an ensemble of classifiers that use features extracted from GSR and from speech does not improve the generalization accuracy significantly.

The rest of this chapter is organized as follows: we review the previous studies on stress detection in speech and physiological signals (Section 2), discuss the benchmark dataset construction, feature engineering from speech and GSR signals, and classification approaches for stress detection (Section 3), summarize the main classification results on the constructed benchmark (Section 4).

2 The problem of stress detection

The concept of stress remains elusive, very broad and used differently in a number of domains. Two models which are commonly used today come from Selye and Lazarus. Selye [17] proposed General Adaptation Syndrome (GAS) model, which identifies various stages of stress response. Selye argues that stress is a disruption of homeostatis by physical (noise, excessive heat or cold) or psychological (extreme emotion, frustration or sleep deprivation) stimuli which alters the internal equilibrium of the body which causes a stress response. Lazarus's model [12] emphasises on two central concepts: appraisal, i.e. individual's evaluation and interpretation of their circumstances, and coping, i.e. individual's efforts to handle and solve the situation. Lazarus argues that neither the stressor nor one's response is sufficient for defining stress, rather it would be one's perception and appraisal of the stressor that determines if it creates stress.

We focus on how the stress shows itself in speech and physiological signals.

Stress detection in speech. The effect of stress in relation to speech production has been well studied over the past decades. Respiration has been found to correlate with certain emotional situations. When an individual experiences a stressful situation, his respiration rate increases. This presumably will increase subglottal pressure during speech, which is known to increase pitch (or fundamental frequency) during voiced section [14]. Moreover, an increase in the respiration rate causes a shorter duration of speech between breaths which, in turn, affects the speech articulation rate.

Another controversial model for detecting stress in speech is Voice Stress Analysis [23], i.e. measures fluctuations in the physiological microtremors present in every muscle in the body, including vocal cords.

Stress detection from physiological signals. Autonomic Nervous Systems (ANS) of a human controls the organs of our body such as the heart, stomach and intestines. ANS can be divided into two divisions, sympathetic and parasympathetic nervous systems. The parasympathetic nervous system is responsible for nourishing, calming the nerves to return to the regular function, healing, and regeneration. On the contrary, the sympathetic nervous system is accountable for activating the glands and organs for defending the body from the threat. The activation of sympathetic nervous system might be accompanied by many bodily reactions, such as an increase in the heart rate, rapid blood flow to the muscle, activation of sweat glands, and increase in the respiration rate.

These physiological changes, such as the electrical activity on the scalp, Blood Pressure (BP), Blood Volume Pulse (BVP), Galvanic Skin Response (GSR) etc., can be measured objectively by using modern technology sensors. These psychological variables can be monitored in non-invasive ways and have been investigated extensively over the past decades.

3 Experimental study

Since there is no publicly available benchmark for studying stress detection from multi-modal sensor measurements we discuss how we constructed one collecting data in controlled settings. Then we discuss the feature extraction from and classification on speech and GSR signals.

3.1 Benchmark construction

Psychological stress elicitation. There have been numerous methods proposed for stress elicitation studies, including the Stroop Color-Word Interference Test [18], the 'Trier Social Stress Test' (TSST) [11] and Trier Mental Challenge Test' (TMCT) that we used for benchmark construction. Stroop found that it took a longer time to read the words printed in a different color than the same words printed in black. This task has been widely utilized as a cognitive stressor able to induce a heightened level of physiological arousal. TSST and TMCT correspond to the public speaking and mental arithmetic stressors that induce considerable changes in the concentration of adrenocorticotropin (ACTH), cortisol, prolactin and heart-rate in independent studies.

Data Collection. During the experiment, several signals were recorded including speech, facial expression, and skin conductance. The speech was sampled at a sampling rate of 44,100 Hz by using two channels. Facial expression was recorded using Handycam Camcorders with High Definition (HD) resolution at $1,440 \times 1,080$ pixels.

To make a GSR sensor measuring the changes in skin conductance we used the LEGO Mindstorms NXT[1] and an RCX wire connector sensor, which converts the analog reading to digital raw values in the range of 0 to 1,023 (Figure 1). The measurements were sampled with 2Hz frequency by using LEGO NXT. Next, the raw value was sent in real time to the computer by means of Bluetooth's connection. We used LEJOS - Java for LEGO Mindstorms[2] open source framework for handling this connection.

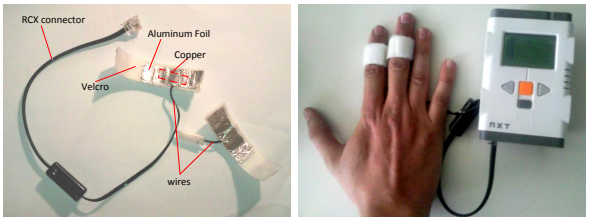

Figure 1. (a) The modified RCX wire connector sensor; (b) GSR device based on LEGO NXT Mindstorms.

We attempted to eliminate the effect of known artifacts into the measurement. Thus, the room temperature where ten subjects participated in the experiment one by one was made constant by means of an air conditioner; no type of physical stressor like noise was present; GSR was collected from the right hand, second phalanx of index and middle fingers. Subjects were of similar age, BMI and were healthy.

[1] http://mindstorms.lego.com/en-us/Default.aspx
[2] http://lejos.sourceforge.net/

The stress experiment lasted for approximately one hour and consisted of three sessions, including baseline (filling the questionnaire and relaxation for 10 minutes), light workload (Stroop-Word congruent color test, an easy mental arithmetic test, and an easy mental subtraction test) and heavy workload triggering stress (Stroop-Word incongruent color test, a hard mental arithmetic test, and a hard mental subtraction test).

3.2 Feature Engineering

Speech features. There are different ways to capture information contained in speech. We considered several subsets including the *smoothed energy* computed using the overlapping time frames (Figure 2), voiced and unvoiced speech (Figure 3), pitch (i.e. the fundamental frequency of the signal), Mel Frequency Cepstral Coefficients (MFCCs) that represent human audio perception (and are the most widely used spectral representation of speech in many applications, including speech and speaker recognition [7] and emotion recognition), and Relative Spectral Transform - Perceptual Linear Perception (RASTA-PLP) [10] features that were shown to be robust to static noise possibly contained in the speech signal.

We used VOICEBOX[3] speech processing toolbox containing the Robust Algorithm for Pitch Tracking (RAPT), proposed by Talkin [19] for computing pitch (and shown to perform better than simple auto-correlation function), and MFCCs extractor and Matlab toolbox for RASTA-PLP[4].

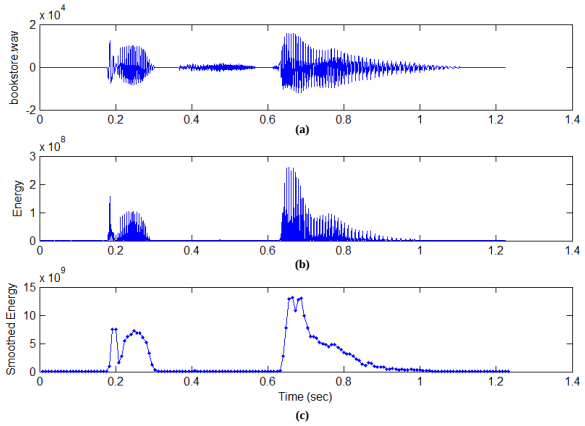

Figure 2. (a) Speech signal (utterance of 'bookstore'), (b) Energy of each sample, and (c) Average energy for each frame.

[3]http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[4]http://labrosa.ee.columbia.edu/matlab/rastamat/

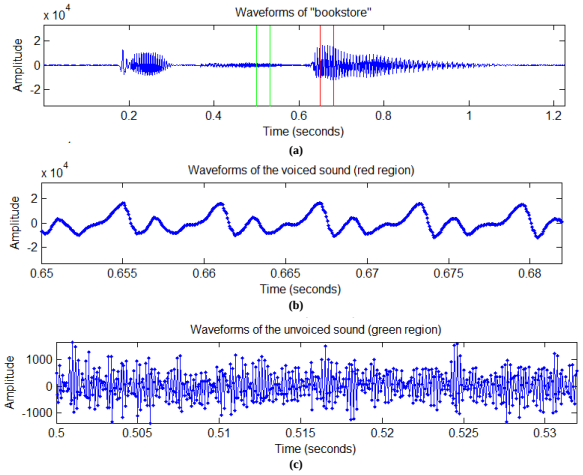

Figure 3. (a) Speech signal of the utterance 'bookstore', (b) Waveform of the voiced sound (red region in (a)), and (c) Waveform of the unvoiced sound (green region in (a)).

GSR features. In general, GSR has a typical startle response (Fig. 4), which is a fast change of the GSR signal in response to a sudden stimulus, characterized by the amplitude and rising time of the signal.

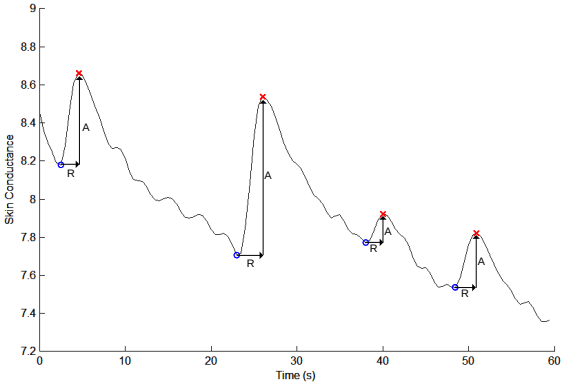

Figure 4. Four GSR startle responses. The peak which is detected by the algorithm is marked with 'x' and the onset is marked with 'o'. The amplitude and rising time of the response are denoted by A and R respectively.

Boucsein has demonstrated that skin conductance is subject to inter-person variability, with differences in age, gender, ethnicity, and hormonal cycle contributing to individual differences [4]. Due to these differences; we normalized the skin conductance signals by subtracting the baseline (values

of the GSR when the user is supposed to be relaxed) minimum and dividing by baseline range [20].

We considered mean, minimum, maximum and standard deviation of skin conductance and peak height; and the total number and the cumulative amplitude, rising time and energy of startle responses in segment. These features were found useful in earlier studies [9].

We performed automatic detection of GSR responses using EDA toolbox [5].

3.3 Classification Methods

We experimented with four different techniques; K-Means classification using vector quantization as a baseline, decision tree classification, Gaussian Mixture Model (GMM) classification [15], which has been shown to work well for speaker recognition, and the Support Vector Machine (LibSVM tool [6]; RBF kernel with cross-validation to choose the best-performing parameters), as a popular general purpose state-of-art classification technique.

We also used logistic regression method for combining the result from two different models (i.e. speech and GSR) into one single regression value. The rationale behind choosing this method is that it is simple, and has been shown to work well for combining two different models in stress detection tasks [13]. Logistic regression requires that the GSR and speech data are available at the same time instance and have been aligned with each other.

We segmented the sensor data streams using a one-minute non-overlapping window. GSR and speech data were aligned manually in offline settings. Then, the instances which only contained the voiced sound of the subject were stored.

We have compared three binary classification tasks: distinguishing recovery vs. workloads (light and heavy), recovery vs. heavy workload and light workload vs. heavy workload. There are 100 instances of recovery class, 110 instances of light workload and 120 instances of heavy workloads (thus there are $110 + 120 = 230$ instances of workload). The corresponding baseline majority class accuracies are $230/340$ or 67.6%, $120/220$ or 54.5%, and $120/230$ or 52.2%.

4 Classification results

First we present subject dependent results (the data from the same subject can be present both in the training and testing sets). Then we also provide the comparative results with 1-subject-leave-out cross-validation to show the corresponding performance for the subject independent case.

[5] https://github.com/mateusjoffily/EDA/wiki

GSR features. Table 1 depicts the average classifier accuracies using 10-times 10-fold cross-validation. GMM, SVM and decision tree method outperformed the baseline (K-Means). The SVM outperforms the other classifiers and can reach accuracy around 80.72% for classifying recovery versus heavy workload session. As expected, differentiating the stress level in the light vs. heavy workload setting is harder than in the recovery vs. heavy workload setting.

Table 1. GSR classification accuracies.

	K-Means	GMM	SVM	DTree
Recov-work	46.1±2.3	70.5±0.5	**79.7±0.8**	73.4±1.3
Recov-heavy	55.5±2.6	74.9±0.8	**80.7±0.6**	77.8±1.3
Light-heavy	53.2±1.0	66.8±0.5	**70.6±1.1**	62.5±1.8

Speech features. We consider here on light vs. heavy workload (110 positive and 110 negative instances) results as the most difficult case. The features which are investigated include pitch, MFCC, MFCC-Pitch and RASTA PLP. A total of 12 pitch features are used, including mean, minimum, maximum, median, standard deviation, range (maximum - minimum) of pitch and its first derivation. We used 144 MFCC features including mean, variance, minimum and maximum of the first 12 cepstral coefficients (excluding the 0-th coefficient), delta coefficients (the first derivative of coefficients) and delta-delta coefficient (the second derivative of coefficients). The feature MFCC-Pitch represents the direct concatenation of MFCC and pitch features. In total, 108 RASTA PLP features were used including mean, variance, minimum and maximum of RASTA PLP coefficients, and first two derivatives. The experimental results using these features with 10-times 10-fold cross-validation are depicted in Table 2.

Table 2. Speech classification accuracies.

	K-Means	GMM	SVM	DTree
Pitch	49.7±2.3	58.8±1.5	**62.1±1.6**	55.6±2.8
MFCC	55.4±1.9	56.8±1.8	**92.4±0.6**	68.9±3.1
MFCC-Pitch	49.2±2.3	59.1±0.9	**92.6±1.6**	70.7±1.3
RASTA PLP	50.6±0.4	52.3±2.8	**91.7±0.9**	71.5±3.0

K-means results in the lowest accuracy and is unsuitable for stress detection. The GMM classifier outperforms K-Means (baseline) with insignificant differences. In general, the GMM accuracies using speech features are lower than using GSR features. In contrast, the SVM method, which is not based on the distribution fit technique, gives a higher accuracy in this setting. This is most likely due to the high dimensionality of speech (e.g. up to 144 dimensions for MFCC) which makes the Gaussian distribution sparse, thus, making the Expectation-Maximization algorithm fail

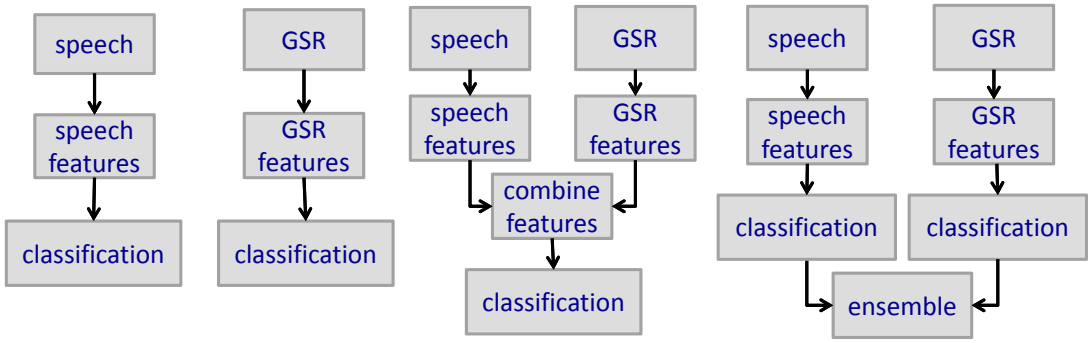

Figure 5. Experimental comparison of four stress classification approaches.

to cluster and fit the data. The decision tree classifier, despite its simplicity, performs relative quite well in this setting. SVMs trained on MFCC and MFCC result in highest accuracy above 92%. Pitch alone is not a good indicator for stress classification.

Fusion of GSR and speech. There are many approaches for combining the GSR and speech. One of them is by using feature enrichment. Both features from speech and GSR are combined together. Afterwards, the classifier is utilized to predict the class output. Another approach is by using ensemble learning such as the logistic regression technique. First, an individual model is built separately. Afterwards, the result of both models is combined is some way. Figure 5 illustrates four stress classification approaches that we compared experimentally. We used logistic regression to combine two individual speech and GSR classifiers. In these experiments we used SVM classifiers as they shown the most promising performance. The corresponding accuracy results are shown in Table 3.

Table 3. Fusion of GSR and speech.

	Features fusion	Models fusion
MFCC and GSR	90.73±1.19	**92.43±0.77**
MFCC-Pitch and GSR	91.34±1.07	**92.47±1.37**
Pitch and GSR	69.04±1.24	**70.17±2.36**

The logistic regression slightly (and not statistically significantly) outperforms feature enrichment yet does not exceed the accuracy of SVM classification based on speech signal alone. Thus, two straightforward methods for combining GSR and speech signals seem not to provide any advantage over individual models. Nevertheless, high values for Cohen's Kappa disagreement statistic computed for the speech and GSR classifiers suggest that GSR and speech signals can complement each other. Thus, methods such as dynamic integration of classifiers that can better utilize the diversity of individual models [21] and that consider evolving settings [22] may give a more promising outcome.

Subject (in)dependent models. The 1-subject-leave-out cross-validation approach was studied to evaluate the model performance for the subject independent case. Figure 6 shows a comparison of 10-times 10-fold cross-validation accuracy results against 1-subject-leave-out cross-validation. It is obvious from this graph that the accuracies of the classifiers drop (dramatically in case of using speech features) when using the subject independent model. Hence, if possible, it is better to address the stress classification problem as a subject dependent model.

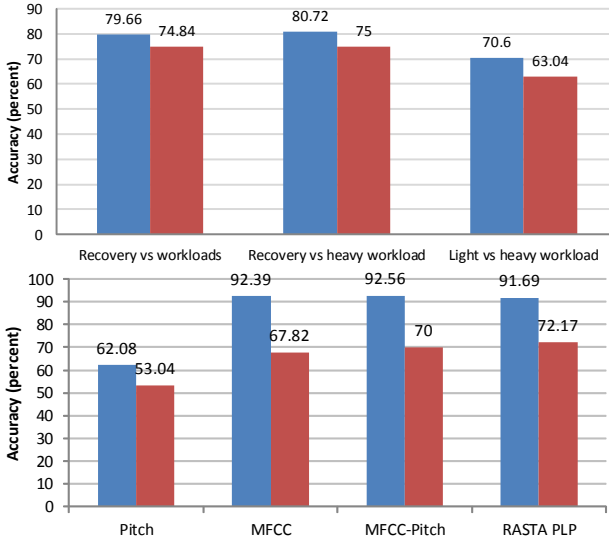

Figure 6. 10x10-fold CV (left bar) and 1-subject-leave-out CV (right bar) accuracies for GSR (top) and speech (bottom) signals.

5 Conclusion

Modern sensor technologies enable the objective measurement of stress. We investigated different models for detecting stress using four classifiers: K-means, GMM, SVM and decision tree. The SVM outperformed the other classifiers on the considered benchmark reaching 92% accuracy by using speech features only. Speech is indeed a good indicator for determining stress in controlled setting. However, our experiments show that we need to have some labeled data from a person for whom we make predictions; otherwise, accuracies may drop substantially.

Accuracies of classifiers trained on GSR features were much lower, at most 70% for differentiating between light and heavy workload. Furthermore, it has been demonstrated in [3] that the GSR signal is varied not only from person to person but also e.g. from one day to another for the same person. Therefore, including other measurements (instead of relying only on GSR-based stress detection) is needed for obtaining more reliable performance. Unfortunately, straightforward approaches for combining both GSR and speech signals using feature enrichment or ensemble learning with logistic regression did not improve the performance significantly. Further research in this direction is needed.

The results presented in this work correspond to the data collected in controlled laboratory settings. Detecting stress levels in the real-life and/or online settings will be a more challenging and more difficult task. There is no guarantee that the signals will be free of noise and artifacts. For instance, the speech may contain background noise or voices of other people, while the GSR signals may contain artifacts owing to the exact placement of the sensor and the physical activity. Also, the speech and GSR may not be available at the same time. Hence one of the directions of further work is in designing mechanisms that can recognize and address all of such practical issues.

References

[1] Y. Ayzenberg, J. H. Rivera, and R. Picard. Feel: frequent eda and event logging – a mobile social interaction stress monitoring system. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 2357–2362, New York, NY, USA, 2012. ACM.

[2] J. Bakker, L. Holenderski, R. Kocielnik, M. Pechenizkiy, and N. Sidorova. Stess@work: From measuring stress to its understanding, prediction and handling with personalized coaching. In *Proceedings of ACM SIGHIT Int. Health Informatics Symposium*, pages 673–678. ACM Press, 2012.

[3] J. Bakker, M. Pechenizkiy, and N. Sidorova. What's your current stress level? Detection of stress patterns from GSR sensor data. In *Proceedings of ICDM Workshops.*, pages 573–580, 2011.

[4] W. Boucsein. *Electrodermal Activity*. The Springer series in behavioral psychophysiology and medicine. Springer, 2011.

[5] J. Cacioppo, L. Tassinary, and G. Berntson. *Handbook of Psychophysiology*. Cambridge University Press, 2000.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[7] S. B. Davis and P. Mermelstein. Readings in speech recognition. chapter Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, pages 65–74. Morgan Kaufmann, 1990.

[8] Y. Deng, D. Hsu, Z. Wu, and C.-H. Chu. Feature selection and combination for stress identification using correlation and diversity. In *12th Int. Symp. on Pervasive Systems, Algorithms and Networks (ISPAN)*, pages 37–43, 2012.

[9] J. Healey and R. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Trans. on Intelligent Transportation Systems*, 6(2):156 – 166, 2005.

[10] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Rasta-plp speech analysis, 1991.

[11] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer. The 'Trier Social Stress Test' – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28:76–81, 1993.

[12] R. Lazarus and S. Folkman. *Stress, Appraisal, and Coping*. Springer Series. Springer Publishing Company, 1984.

[13] I. Lefter, L. J. M. Rothkrantz, D. A. van Leeuwen, and P. Wiggers. Automatic stress detection in emergency (telephone) calls. *IJIDSS*, 4(2):148–168, 2011.

[14] P. Rajasekaran, G. Doddington, and J. Picone. Recognition of speech under stress and in noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, volume 11, pages 733 – 736, apr 1986.

[15] D. A. Reynolds. Gaussian mixture models. pages 659–663, 2009.

[16] J. H. Rivera, R. Morris, and R. Picard. Call center stress recognition with person-specific models. In *Affective Computing and Intelligent Interaction*, volume 6974 of *Lecture Notes in Computer Science*, pages 125–134. Springer, 2011.

[17] H. Selye. Stress and The General Adaptation Syndrome. *The British Medical Journal*, 1(4667), June 1950.

[18] J. Stroop. Interference in serial verbal reactions. *J. Exp. Psychol*, 18:643 – 661, 1935.

[19] D. Talkin. A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, 495:495–518, 1995.

[20] L. G. Tassinary. Inferring psychological significance from physiological signals. *American Psychologist*, 45:16–28, 1990.

[21] A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Diversity in search strategies for ensemble feature selection. *Information Fusion*, 6(1):83–98, 2005.

[22] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen. Dynamic integration of classifiers for handling concept drift. *Information Fusion, Special Issue on Applications of Ensemble Methods*, 9(1):56–68, 2008.

[23] J. Z. Zhang, N. Mbitiru, P. C. Tay, and R. D. Adams. Analysis of stress in speech using adaptive empirical mode decomposition. In *Proc. of 43rd Asilomar Conference on Signals, Systems and Computers*, pages 361–365. IEEE Press, 2009.