Call Center Stress Recognition with Person-Specific Models

Javier Hernandez, Rob R. Morris, and Rosalind W. Picard
Media Lab, Massachusetts Institute of Technology, Cambridge, USA
{javierhr,rmorris,picard}@media.mit.edu

Abstract. Nine call center employees wore a skin conductance sensor on the wrist for a week at work and reported stress levels of each call. Although everyone had the same job profile, we found large differences in how individuals reported stress levels, with similarity from day to day within the same participant, but large differences across the participants. We examined two ways to address the individual differences to automatically recognize classes of stressful/non-stressful calls, namely modifying the loss function of Support Vector Machines (SVMs) to adapt to the varying priors, and giving more importance to training samples from the most similar people in terms of their skin conductance lability. We tested the methods on 1500 calls and achieved an accuracy across participants of 78.03% when trained and tested on different days from the same person, and of 73.41% when trained and tested on different people using the proposed adaptations to SVMs.

Keywords: Stress recognition, skin conductance, interpersonal variability, Support Vector Machines, Affective Computing.

1 Introduction

Chronic psychological stress carries a wide array of pathophysiological risks, including cardiovascular disease, cerebrovascular disease, diabetes, and immune deficiencies [8]. An important step in managing stress, before it becomes chronic, is recognizing precisely when and where it occurs. Technologies that automatically recognize stress can be extremely powerful, both diagnostically and therapeutically. As a diagnostic tool, technologies such as these could help individuals and clinicians gain insight into the conditions that consistently provoke maladaptive stress responses. As a therapeutic tool, these technologies could be used to automatically initiate stress-reduction interventions. In stressful work settings, such as a call center, these technologies could not only lead to more timely and reduced-cost interventions, but also to more productive environments where employees could better manage their workload, so that they could provide a better experience for customers.

While research on automated stress recognition has taken many different forms, the systems that have been proposed in the engineering literature typically contain two principle components: 1) a sensor-based architecture that records

S. D´Mello et al. (Eds.): ACII 2011, Part I, LNCS 6974, pp. 125–134, 2011.

relevant features and 2) a software-based system that makes predictions about an individual's current stress level. The sensing modalities can take many forms, including audio and visual modalities, but biosensors provide the most direct access into the physiological changes that accompany stress-induced changes [3].

While great strides have been made in real-life biosensing [13], the computational task of inferring stress levels from biosensor data is still a considerable challenge. There is often great variability in how people experience stress [10] and how they express it physiologically [11], and this interpersonal variability can stymie efforts to build a one-size-fits all stress recognition system. This work explores using data from each individual to help manage the problem of interpersonal variability. In particular, we modify the loss function of SVMs to encode a person's tendency to report stressful events, and give more importance to the training samples of the most similar participants. These changes were validated in a case study where skin conductance (SC) was monitored in nine call center employees during a one-week period of their regular work.

This paper is organized as follows. Section 2 reviews previous studies on the subject of this work. Section 3 provides details about the data collection. Section 4 presents the problem of interpersonal variability and proposes two complementary methods to address it. Section 5 explains the data preprocessing and experimental protocols. Section 6 provides results and analysis.

2 Background and Previous Work

2.1 Physiological Stress and Skin Conductance

Stress-induced changes can be monitored with biosensors, and a particular focus is often placed on the sympathetic nervous system, which is designed to mobilize the body's resources in response to a challenge or a threat. While most visceral organs are dually innervated by both the para- and sympathetic nervous systems, the eccrine sweat glands are thought to be solely controlled by the sympathetic nervous system [3]. Thus, skin conductance sensors that measure eccrine sweat gland activity are often used to monitor sympathetic nervous system activity.

A century of short-term lab measurements have shown that SC is subject to inter-person variability, with differences in age, gender, ethnicity, and hormonal cycles contributing to individual differences [3]. Furthermore, many researchers suggest that stable personality differences may contribute to differences in skin conductance lability - a psychophysiological trait characterized by high SC responsivity and slow habituation [12]. As early as 1950, researchers have seen links between SC lability and such personality characteristics as emotional expressiveness, and antagonism [5]. Moreover, individuals defined as SC labiles have been seen to show greater myocardial reactivity in response to stress [9]. When developing stress recognition algorithms that incorporate measures of SC, interpersonal sources of variance should be considered.

2.2 Automatic Stress Recognition

Several automatic stress recognition techniques have been explored in the research literature. In most cases, data are collected in the laboratory where variables that introduce noise are controlled or eliminated.

Researchers have explored a variety of classification methods, and techniques to reduce interpersonal variability. Barreto, Zhai and Adjouadi [1], for example, used SVMs to discriminate between stressful and non stressful responses in a laboratory setting. The SVMs outperformed other classification algorithms, obtaining an accuracy of 90.1%. Various physiological signals were used in the classification, including SC, blood volume pulse, pupil diameter (PD) and skin temperature (ST). To account for participant variability, they divided extracted features from each participant with their corresponding baseline features. In a separate study, Setz et. al. [14] used SC to automatically distinguish between cognitive load and psychosocial stress. In this case, Linear Discriminant Analysis (LDA) obtained 82.8% accuracy, outperforming SVMs. Setz et al. found that the average number of SC peaks, as well as their height distributions, were the most relevant features to the problem. To account for participant variability, distributions were computed for each participant independently. In another study, Shi et al. [15] discriminated between stressful and non-stressful responses under social, cognitive and physical stressors. They obtained 68% precision and 80% recall using SVMs with SC, electrocardiogram (ECG), respiration (R) and ST. The problem of participant variability was addressed by subtracting a person-specific parameter to the features of each participant. This parameter was estimated as the average feature of all-non-stressful events of the participant.

In an effort to automatically recognize stress in a real-life setting, Healey and Picard [6] monitored ECG, electromyogram, SC and R from people during a driving task. They used LDA to automatically discriminate between low (at rest), medium (highways) and high (city) levels of stress with 97% accuracy. In this case, the signals from each participant were normalized between zero and one, as proposed by [11].

All of these studies, except for Setz et. al. [14], used a combination of physiological signals, an approach that typically improves recognition accuracy. Nevertheless, some of the signals, such as PD and ECG, may not be easily recorded in real-life settings where comfortable and inconspicuous sensors are required to preserve natural behavior.

3 Study Design

Location and Participants. The study was conducted at a call center in Rhode Island, and was approved by the Institutional Review Board at the Massachusetts Institute of Technology. Nine call center employees (five females and four males) agreed to participate in the study. The employees all had the same job description and they all handled the same types of calls.

Throughout the course of one week, and only during work hours, participants wore a wristband biosensor and made self-report ratings at the end of each call

they received. Besides those two, minimally invasive conditions, the participants went about their work as usual. Their day is primarily spent on the phone, and they handle high volumes of calls, many of which come from angry and frustrated customers.

Data Collection. Three sources of data were collected in this study: SC, self-report measures, and worker call logs. SC was collected at a sampling rate of 8 Hz, similar to [10] and [15], and was recorded from dry Ag-AgCl 1cm diameter electrodes on the wrist, using an early beta version of the Affectiva[1] Q^{TM}Sensor, a commercial sensor based on [13].

Throughout the study, participants were also asked to rate each call they received in terms of stress. Specifically, they were asked "How was the last call?" using a 7 point likert scale, with the endpoints labeled as "extremely good" indicating non-stressful and "extremely bad" indicating very stressful. While this question may not capture other types of stressors, it allowed for quick (1-2 seconds) and non-disruptive self-report ratings. The call center also provided break times and detailed call logs for each participant containing the start-time, end-time, and duration of every call our participants received.

A total of 1500 calls were included in our study, averaging 4.51 minutes in length. Calls that had missing stress ratings, or corrupted SC (due to beta hardware problems or motion artifacts), were excluded. Fig. 1 shows a one day example of collected raw data.

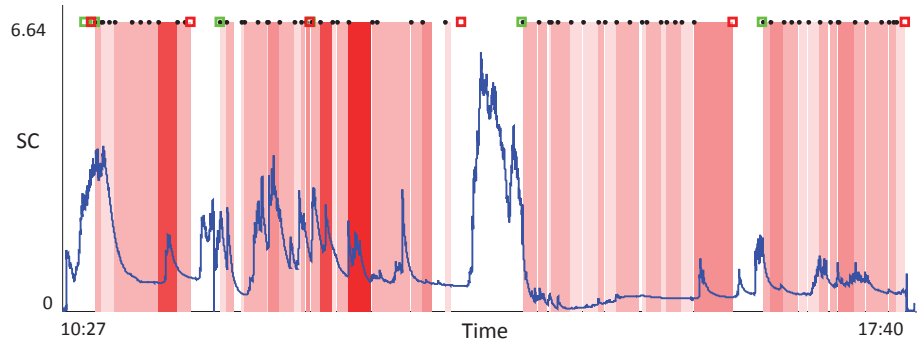

Fig. 1. Example of data from one participant that contain calls (*dots*), stress ratings (*darker areas* represent more stressful calls), and break times (*squares*)

4 Proposed Method

Throughout this paper, we shall focus on the problem of supervised classification. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ be an i.i.d. training set, where $\mathbf{x}_i$ represents the feature vector of the sample i, and y_i its class label, where $y_i = \{-1, 1\}$. Let the class priors of this set be $P_+ = \frac{\#y=1}{n} = \frac{n_+}{n}$ and $P_- = \frac{n_-}{n}$. Similarly, we define the testing set as $\{(\overline{\mathbf{x}_i}, \overline{y_i})\}_{i=1}^{\overline{n}}$, and its priors $\overline{P_+}$, and $\overline{P_-}$.

[1] http://www.affectiva.com

We consider the problem where training data comes from the observation of a set of participants, and the testing data belongs to a new participant. This methodology introduces the common problem of participant variability, which usually violates the i.i.d. assumption and leads to an overall decrease in performance. To address the participant variability problem, we propose incorporating information of the testing participant into the loss function of SVMs.

Support Vector Machines [2] are considered state-of-the-art supervised classification algorithms, and their main goal is to find the hyperplane $\mathbf{w}$ that maximizes the margin between data samples belonging to two classes (e.g., stressful vs non-stressful responses). The standard formulation of SVMs is as follows:

$$\min_{\mathbf{w}} \quad \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{regularization} + \underbrace{\frac{C}{n}\left(\sum_{i\in\{y=+1\}}^{n_+}\xi_i + \sum_{j\in\{y=-1\}}^{n_-}\xi_j\right)}_{loss\ function}, \tag{1}$$

$$s.t.\ \ y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1-\xi_i \quad and \quad \xi_i \geq 0, \quad i = 1, 2, \ldots n \tag{2}$$

where C is the misclassification cost, and ξ_i is the slack variable for the sample $\mathbf{x}_i$. For any new sample $\overline{\mathbf{x}}$, prediction is performed through $\overline{y} = \mathbf{w}^T\overline{\mathbf{x}}$.

4.1 Changing Class Priors

In the context of stress recognition, class priors indicate the probability to report stressful events. In equation 1, priors of the training data are directly integrated into the loss function, and will condition the predictions of the classifier. Since different people may report more or less stressful events, we propose modifying SVMs' loss function to encode the class priors of the testing participant.

A standard method to modify the class priors is the introduction of class weights (S_+ and S_-) for each type of misclassification error such as:

$$loss\ function\ = \frac{C}{n}\left(\sum_{i\in\{y=+1\}}^{n_+} S_+\xi_i + \sum_{j\in\{y=-1\}}^{n_-} S_-\xi_j\right). \tag{3}$$

If $\frac{S_+}{S_-} = \frac{P_-}{P_+}$, the classifier will tend to equally predict positive and negative samples [7]. To predict with the same priors of the testing data, we propose to use $S_+ = \frac{\overline{P_+}}{P_+}$, and $S_- = \frac{\overline{P_-}}{P_-}$. These weights come from enforcing the testing class priors

$$\overline{P_+} = \frac{n_+S_+}{n_-S_- + n_+S_+} \quad and \quad \overline{P_-} = \frac{n_-S_-}{n_-S_- + n_+S_+}, \tag{4}$$

while preserving the same magnitude of the misclassification error:

$$n_+ + n_- = S_+n_+ + S_-n_-. \tag{5}$$

4.2 Selecting Training Samples

As described in Section 2, most of the approaches to address the interpersonal variability problem are based on feature transformations. Although these normalizations work well in practice, some participants may be less relevant than others to the classification because their display of physiologically responses is very different to the ones of the testing participant. Using a small set of unlabeled testing data, we propose finding the similarity of each training subject with the testing subject and use it during classification. We can encode this information as follows:

$$loss\ function\ = \frac{C}{n}\left(\sum_{p=1}^{r} v_p \sum_{i \in participant_p}^{n^p} \xi_i\right)\ , \tag{6}$$

where r is the number of training participants, n^p is the number of samples of the participant p, and v_p defines the similarity of the participant p for classification, based on SC lability. In particular, we computed the average number of peaks (at least 0.05 μS of amplitude) per second and their height average for each training participant, and used k-Means clustering with $k = 2$ to divide the participants. Given a new testing participant, we computed the same information and assigned $v = 1$ to the participants of the closer cluster, and $v = 0$ to the participants of the furthest one.

5 Experimental Setting

Preprocessing. Prior to our analysis, stress ratings were normalized for each participant in order to use all of the scale and to attenuate subjectivity. Furthermore, since the call ratings were quite unbalanced (see Table 1), we transformed the problem to a binary case where calls defined as definitely non-stressful (rating of "extremely good") were grouped into the negative class, and the remaining calls were grouped into the positive class. Table 2 shows the average P_+ value of different days for each participant. As hypothesized, the tendency to report stressful events is very different between participants and similar for different days of the same person.

Table 1. Distribution of call ratings (1 - "extremely good" and 7 - "extremely bad")

Rating	1	2	3	4	5	6	7
Number of Calls	657	379	163	139	45	83	34

Exponential smoothing ($\alpha = 0.8$) was applied to the SC signals to reduce noise and motion artifacts. Skin conductance signals for each participant were also normalized between zero and one to reduce the overall variability of the group [11]. From each signal, we extracted the following features: duration, maximum and minimum values and their relative positions to the signal duration, mean, standard deviation, slope between the first and last signal values, number of zero crossings, and quantile thresholds to capture the distribution of peak heights as described in [14]. These features were normalized to have zero mean and unit standard deviation.

Table 2. Average and standard deviation (STD) of P_+ for the nine participants

Participant	1	2	3	4	5	6	7	8	9	Average	STD
Average (%)	97.06	86.63	78.88	75.51	66.35	56.77	34.76	28.73	10.93	59.51	29.05
STD (%)	2.69	10.89	10.06	14.80	5.45	11.68	2.93	31.01	2.86	10.26	8.96

Experiments. Two testing protocols were used for the analysis. The first protocol (A) used leave-one-day-out cross-validation to obtain the stress ratings of one participant. That is, we used all days of a participant's data to train the algorithm, except one day that was used for testing. The process was repeated until all days were used as testing data. We expect this protocol to give the best performance for this data set, because both training and testing data come from the same participant. In practice, however, this protocol scales badly because it requires annotated information for each new participant. The second protocol (B) used leave-one-participant-out cross-validation. Here, the algorithm was trained with data from eight participants to predict the stress levels of the remaining participant, and it was repeated until all of the participants had been part of the testing data. This is a more realistic but difficult protocol in which the distribution of the training data and the testing data are dissimilar due to interpersonal variance. We tested the proposed modifications in this protocol with the expectation that it would mitigate the variance while preserving scalability.

To perform classification, we used the publicly available LIBSVM library [4] that provides an efficient implementation of SVMs. We used the Radial Basis function as the kernel function to allow non-linear decision boundaries. For each training set, leave-one-participant/day-out was also used to find the parameters ($\log_2 C \in \{-3:2:5\}$) and $\log_2$ *kernel width* $\in \{-15:2:-1\}$) that maximized the following expression:

$$\frac{TN}{2(FN+FP+2TN)} + \frac{TP}{2(FN+FP+2TP)}\ , \tag{7}$$

where TP and TN are the number of correctly predicted stressful (true positives) and non-stressful (true negatives) calls respectively, and FN and FP correspond to the number of misclassified stressful (false negatives) and non-stressful (false positives) calls respectively. This expression enforces the same relevance to both classes independently of their class priors.

6 Results

Following the previous experimental settings, Fig. 2 shows the results for protocol A, protocol B, and improvements of protocol B - correcting class priors (CCP) and selecting training samples (STS). As expected, when no improvements were applied, protocol B showed consistently lower average performance than protocol A, 58.45% and 78.03% respectively. This finding confirms that participant variability is difficult to model even though our data was appropriately

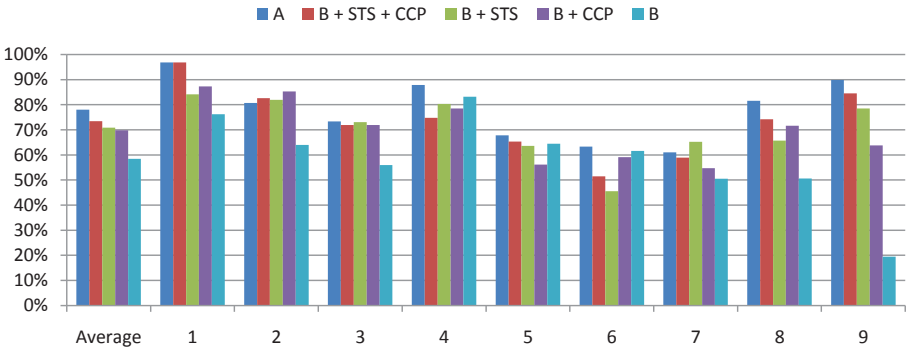

Fig. 2. Classification accuracy for each participant

normalized for all experiments. While CCP and STS individually increased the average accuracy to 69.83% and 70.91% respectively, the combination of the two improvements increased performance to 73.41%. Moreover, STS has the additional benefit of reducing the amount of training samples and therefore reducing the computational cost of the training phase. Closer inspection of Fig. 2 shows that the improvements did not increase performance for two out of the nine participants (4 and 6). No significant relationships could be made between the two participants, but a replication of similar experiments with a larger number of participants could shed light on this topic. To compare the overall performance, Fig. 3 (left) shows the Receiver Operating Characteristic (ROC) curves of protocol A, protocol B and B + STS + CCP. By observing the area under the curve (AUC), we can conclude that both improvements increased the overall accuracy.

Although accuracy has been used for most of the research papers to compare performance, it may not be the most adequate metric for real-life settings where class labels may be very unbalanced. For instance, accuracy values could be high if the algorithm predicted just the most likely class which could potentially

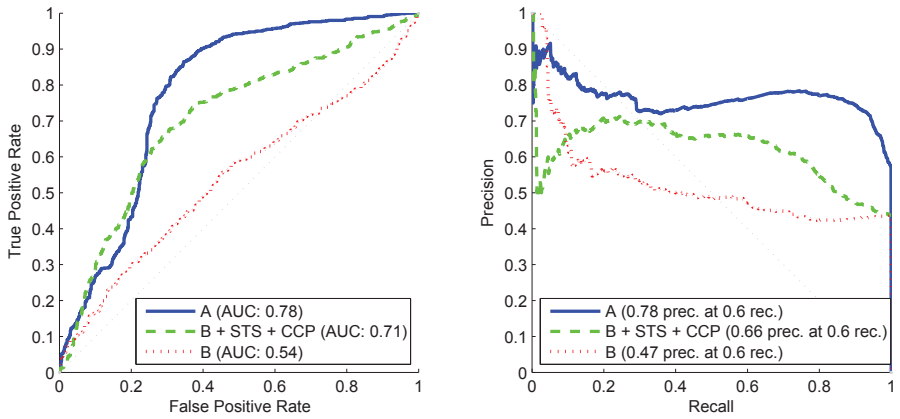

Fig. 3. ROCs and precision-recall curves

ignore the class of interest (e.g., stressful calls.) As a complementary metric, we use precision-recall curves (see Fig. 3, right.) To analyze this curve, we can study a real case application where the company wants to collect stressful calls to train their new employees. In this case, the company wants to know how many of the calls predicted as stressful by the classifier were also reported as stressful by the employees. For instance, if we optimize our methods to correctly detect stressful calls 60% of the time (i.e., recall = 0.6), the percentage of these detections that are also reported as stressful calls (precision) is 78.40% for protocol A, 65.84% for B + STS + CCP, and 46.82% for B alone. These results are in line with the results using accuracy and, therefore, we can conclude that the proposed methods partly address the participant variability problem.

7 Conclusions

This is one of the few research studies on stress recognition in an uncontrolled (real-life) setting. Unlike many other studies on workplace stress, we did not alter the working conditions to artificially create stressful scenarios. This naturalistic approach introduced undesired real-life variables (e.g., unbalanced reports, artifacts), many of which accentuated the problem of participant variability. In this context, we proposed two methods to account for individual differences in order to discriminate stressful vs. non-stressful calls of nine call-center employees.

The two improvements - correction of the class priors and the selection of training samples - rely on the use of data from the testing participant. In many cases, the recovery of testing class priors may be unfeasible or expensive but, in this case, simple questionnaires can be used to obtain that information. As we showed great similarity in participants' stress reports across days, we can also use one day of labeled monitoring to obtain the priors. As for the STS, we explored the use of SC lability to encode similarity between participants, a method that does not require any labeling. In the future, we intend to explore other similarity measures based on demographic characteristics (e.g., age, gender or ethnicity), and we intend to incorporate temporal models (e.g., Hidden Markov Models) to capture the dynamics of stress.

In this paper we have illustrated the benefits of using person-specific models for stress recognition in a call center setting, but the methods explored in this paper can generalize to many areas of Affective Computing. Indeed, participant variability is a common issue in many types of affect recognition applications, and new methods are sorely needed to help tackle this problem.

Acknowledgements. This work was supported in part by the MIT Media Lab Consortium. Javier Hernandez was supported by the Caja Madrid fellowship.

References

1. Barreto, A., Zhai, J., Adjouadi, M.: Non-intrusive physiological monitoring for automated stress detection in human-computer interaction. In: ICCV-HCI, pp. 29–38 (2007)

2. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: 5th Annual ACM workshop on Computational Learning Theory, pp. 144–152. ACM Press, New York (1992)
3. Boucsein, W.: Electrodermal Activity. Plenum Press, New York (1992)
4. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), software `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
5. Crider, A.: Personality and electrodermal response lability: an interpretation. Applied Psychophysiol Biofeedback 33(3), 141–148 (2008)
6. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans. Intell. Transport. Syst. 6, 156–166 (2005)
7. Huang, Y.M., Du, S.X.: Weighted support vector machine for classification with uneven training class sizes. In: 4th International Conference on Machine Learning and Cybernetics, vol. 7, pp. 4365–4369. IEEE Press, Los Alamitos (2005)
8. Cacioppo, J.T., Tassinary, L.G., Berntson, G.G.: Handbook of Psychophysiology. Cambridge University Press, Cambridge (2000)
9. Kelsey, R.M.: Electrodermal lability and myocardial reactivity to stress. Psychophysiology 28(6), 619–631 (1991)
10. Lunn, D., Harper, S.: Using galvanic skin response measures to identify areas of frustration for older web 2.0 users. In: International Cross Disciplinary Conference on Web Accessibility, p. 34. ACM, New York (2010)
11. Lykken, D.T., Venables, P.H.: Direct measurement of skin conductance: A proposal for standarization. Psychophysiology 8(5), 656–672 (1971)
12. Mundy-Castle, A.C., McKiever, B.L.: The psychophysiological significance of the galvanic skin response. Experimental Psychology 46(1), 15–24 (1953)
13. Poh, M., Swenson, N., Picard, R.: A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Trans. Biomed. Eng. 57(5), 1243–1252 (2010)
14. Setz, C., Arnrich, B., Schumm, J., La Marca, R., Troster, G., Ehlert, U.: Discriminating stress from cognitive load using a wearable eda device. IEEE Transactions on Information Technology in Biomedicine 14(2), 410–417 (2010)
15. Shi, Y., Nguyen, M.H., Blitz, P., French, B., Fisk, S., De la Torre, F., Smailagic, A., Siewiorek, D.P., al' Absi, M., Ertin, E., Kamarck, T., Kumar, S.: Personalized stress detection from physiological measurements. In: International Symposium on Quality of Life Technology (2010)