EEG Signal Classification in Usability Experiments

Vagner do Amaral, Leonardo A. Ferreira, Plinio T. Aquino Júnior and Maria Claudia F. de Castro
Department of Electrical Engineering, Centro Universitário da FEI
São Bernardo do Campo, São Paulo, Brazil
e-mail:{vamaral, laferreira, plinio.aquino, mclaudia}@fei.edu.br

***Abstract*—The affective computing aims to detect emotional states during the interaction between the user and the machine allowing the use of this information in decision-making processes. EEG signals related to emotional states can be applied to the context of software usability providing more resources to the validation process and the identification of the degree of user satisfaction. This work aims to establish a relationship between EEG signals and the user opinion about the usability of some Facebook privacy features. Based on the assumption that there are variation in brain activity during the execution of tasks labeled as "easy" or "difficult", a performance evaluation was done based on a Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) classifiers. The Mean Power Spectral Density, in 7 frequency bands, from 8 electrodes in F, C, P, and O areas were used as features. The classification rates showed a small advantage of the SVM when all the 28 variables were used. However, when the 13 variables pointed by the Mann-Whitney U test were used, LDA showed good discrimination capability. The electrodes in F and C areas, related with cognition and motor functions, rejected null hypothesis in almost all frequency bands during the execution of the tasks, showing that it is possible to recognize the studied emotional states. Despite the fact that this was a preliminary study, it showed the feasibility of using the EEG as a potential source of information to be added to software usability testing.**

***Index Terms*—Usability Testing, EEG Signal, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Pattern Recognition.**

I. Introduction

Emotion recognition is a simple task among humans and plays an important role in interpersonal relationships. However, human-machine interface systems still have a diminished capability to interpret and deal with them. The affective computing aims to fill this gap by detecting emotional states during the interaction between the user and the machine, allowing the use of this information in decision-making processes. The works [1], [2], [3], [4] and [5] showed the possibility of applying neural signals in this process, through the association between electroencephalogram (EEG) and emotional states.

Murugappan et al. [1], reported the application of two well known classifiers, K Nearest Neighboar (KNN) and Linear Discriminat Analysis (LDA), to categorize human emotions stimulated by video clips using spatial filtering and wavelet transform to extracting statistical features from EEG signals. This study showed a good average classification rate in some of the proposed approaches.

Kimura et al. [2], described a study toward constructing a EEG measurement method for usability evaluation. They conducted an experiment to reveal the optimal time length of EEGs for human-computer interaction tests by analyzing changes of signal features over time. The results show that the accuracy of usability evaluation can be improved by using length of time window about 56 seconds.

Lee and Seo [3], applied a variety of biological signals such as Electroencephalograph (EEG), Electrocardiogram (ECG) and Electromyogram (EMG) to evaluate users emotional reaction to different web interface designs and conducted a typical usability test in the same environment to compare these two methods. They concluded that usability testing using biosignals is a good method to use for web evaluation.

Masaki et al. [4], tried to evaluate the human-computer interaction experiences quantitatively using EEG. They conducted experiments to observe the relationships between EEG and usability experience in using software. From findings, they concluded that EEG signals can be an indicator for measuring the software use experiences quantitatively.

Koelstra et al. [5], presented a physiological multimodal database and explored the research on affective states in humans stimulated by video clips. This study describes in details the acquisition process settings and reports an extensive statistical analysis between EEG signals and the self-assessment volunteers questionnaire. The results indicated a significant correlation between these data.

Therefore, based on those results, it seems that the procedure of using EEG related to emotional states can be applied to the context of software usability, providing more resources to the researchers aiming at validating user opinions reported during a trial to identify the degree of user satisfaction of a particular system.

In this context, this work purposes a study to establish a relationship between EEG signals and the user opinion about the usability of some Facebook privacy features. Based on the assumption that there are variation in brain activity during the execution of tasks considered by the user as "easy" or "difficult", a performance evaluation is done based on two classifiers methods, a Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM). The Mean Power Spectral Density, in 7 frequency bands, from 8 electrodes in F, C, P, and O areas are used as features.

II. Usability testing

Usability Testing or Usability Experiments involves observing real users performing real-life or simulating tasks with their interactive system, as they would during normal usage scenarios. The usability researcher supervises all tests to record user's misinterpretations, errors, stumbling blocks, expectations and attitudes. The observations are then analyzed, problems are described along with high effect solutions. It is recommended conducting usability experiment as it eliminates presumption and get honest, first-hand feedback from real users. The central benefit is that the development group gets to detect problems earlier to save redundant period and costs of interface redesign later. The final product reduces burden and cost of post-launch support requests. Also gain robust brand loyalty and an unmatched competitive benefit [6].

The number of users as well as the number of tasks to be supported out by them depend on the complexity of the specific product. Tests can be conducted in usability labs, but to capture the usage situations more realistically, they should be done int the user's natural environment.

Usability tests can be carried out at any stage of product development. For best results, tests should be conducted from the conceptual stages throughout the interface systems development process, till the interface system is ready to launch. Early tests are conducted with "prototypes". A prototype can be as basic as a paper sketch or a semi-finished system with limited functionality. Testing with prototypes is fast, low-cost and effective, and it helps the detection of usability faults early and save redundant redesign budgets. At this point, users should be characterized. Several techniques are available: user profile, user roles, personas [7], [8].

However, usability experiment can be applied in systems published to the general public. In this case, usability tests are used to define improvements in product application.

III. Linear Discriminant Analysis

Initially proposed by Fisher [9], it is a well know method to linearly classify a dataset from its many features. The Linear Discriminant Analysis (LDA) classifies a dataset based on the relation between the dispersions within classes and between classes, in order to find the dimension that best classify a dataset in a linear way [10].

From the between classes dispersion S_b, given by the equation:

$$S_b = \sum_c (\mu_c - \mu)(\mu_c - \mu)^T, \tag{1}$$

in which, c are the classes, μ is the mean between every data points and μ_c is the mean within the classes and the within classes dispersion S_w is calculated with:

$$S_w = \sum_c \sum_{j \in c} p_c \, (x_j - \mu_c)(x_j - \mu_c)^T, \tag{2}$$

being, x the value of the data in the dataset, j the data classified as belonging to the class c and p_c is the probability that these data belongs to the class c, given by the relation of the number of data points in the class and the total in every class, the goal of the LDA is to obtain the matrix P_{lda}, that maximizes the relation between S_b and S_w, that is

$$P_{lda} = \arg\max_P \frac{|P^T S_b P|}{|P^T S_w P|}, \tag{3}$$

which will provide the hyperplane that should be used in order to best classify the data using only one available dimension. The LDA gives the best dimension that describes a dataset dispersion given it's features and therefore, is able to analyze the data by reducing it's dimensionality.

IV. Support Vector Machines

Another way to linearly classify a dataset is by using the Support Vector Machines (SVM) proposed by Vapnik [11]. The SVM uses the concept of optimal hyperplane, an hyperplane that separates the data without error with the maximal distance between the data vector and the plane. While the dataset cannot always be separated by linear functions, the dimensionality increase with kernel functions is a common approach by using the SVM classifier.

Vapnik's [11] proposal is grounded in the transformation of a linearly inseparable dataset in a new one that is linearly classifiable, increasing the number of dimensions using only the data from the original dataset. After this data transformation, it becomes possible to classify the original dataset using those new dimensions generated by the transformation [10].

In order to understand how the SVM classify a given dataset, first it must be explained how it estimates the hyperplane and the concept of kernel functions used to increase the dimensionality.

A. Optimal hyperplane or maximal margin hyperplane

As defined by Vapnik [11], a dataset of features $X \in \Re^n$ and output $y \in \{-1, 1\}$ with m examples and described as:

$$\forall X \in \Re^n, y \in \{-1, 1\}, (X_1, y_1), \ldots, (X_m, y_m), \tag{4}$$

can be separated by an hyperplane in the form of a linear function:

$$(w \cdot X) - b = 0 \tag{5}$$

and if this computed hyperplane classify the dataset without error and has the maximum distance between the vector and the hyperplane, considering all other possible vectors, this computed hyperplane is called optimal or maximal margin hyperplane.

A way to describe a separating hyperplane between two classes $y \in \{-1, 1\}$ is given by the inequalities:

$$\begin{cases} (w \cdot X_i) - b \geq 1 & \text{if } y_i = 1, \\ (w \cdot X_i) - b \leq -1 & \text{if } y_i = -1 \end{cases} \tag{6}$$

and the maximal margin hyperplane can be calculated by minimizing the function:

$$\Phi(w) = \frac{1}{2}(w \cdot w) \tag{7}$$

In a dataset that can be linearly separable, the concept of maximal margin hyperplane is enough to classify the dataset without errors and the Support Vector Machine can be used without the necessary increase of dimensionality. While for a non linearly separable dataset the features X are mapped to a high-dimension feature space using a nonlinear function chosen a priori and the optimal hyperplane is then constructed in this new high-dimension space.

B. Kernel functions and the increase of dimensionality

There are several ways to increase the dimensionality of a dataset, some of those are more commonly used as Kernel functions. Usually, a Kernel function $K(x, x')$ computes the similarity between two features x and x' in a Hilbert Space, being the three most common the Linear, Polynomial and RBF Kernels. By using any of these three functions it is possible to increase the dimensionality of a given dataset so that it might become linearly separable by using a SVM classifier.

The increase of dimensionality by itself presents a problem in the computation of the new dataset and the resources need for such calculus. Nevertheless, as shown by Vapnik in the same work as the SVM is proposed [11], it is possible to use only the inner products of the support vectors and the dataset vectors, although the original dataset dimensionality might already be enough to make the computation take a long time or even become impossible to analyze.

V. Materials and Methods

This paper considers usability experiment on the finished interface. For the definition of interface improvements it is important to map the user's behavior and cognition. For this purpose the use of EEG can be an alternative. Figure 1 shows a data analysis workflow adopted in this study and the following sections describe the methods employed in this work to integrate the fields of usability, signal processing and statistic.

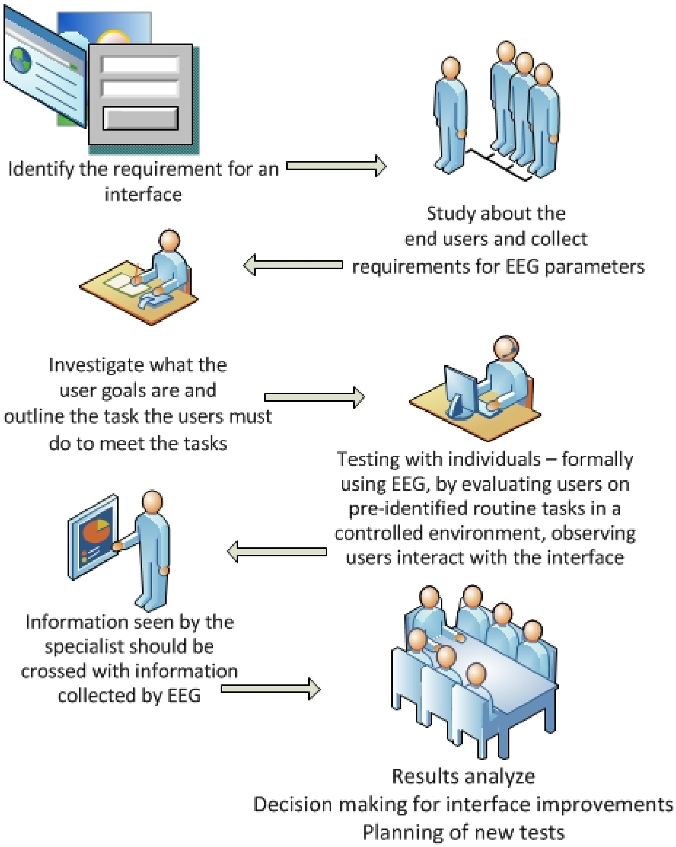
Identify the requirement for an interface
Study about the end users and collect requirements for EEG parameters
Investigate what the user goals are and outline the task the users must do to meet the tasks
Testing with individuals – formally using EEG, by evaluating users on pre-identified routine tasks in a controlled environment, observing users interact with the interface
Information seen by the specialist should be crossed with information collected by EEG
Results analyze
Decision making for interface improvements
Planning of new tests

Fig. 1. Data analysis workflow.

A. User Testing Setup

This study applied usability experiment in FACEBOOK interface. We considered two use contexts: iPAD and desktop computer use. The goal was to evaluate user behavior in these two use contexts. The traditional experiments usability not provides detailed information of the user's cognitive profile. Only enables a user through observation by the specialist. It is a subjective result, because it depends on the interpretation of the expert who observes and judges the experiment.

The user received information about the function and importance of the test. The test environment has special layout, composed environments separated by a mirror. The Morae software was used to document the test. Morae [12] comes equipped with real-time chat for communicating during testing, enhanced pluggable architecture for a customizable testing experience, searchable table of contents for easy sharing. This software does not provide support to EEG, but increases the productivity of collection data and analysis in traditional usability testing (without EEG).

B. EEG: Aquisition and Preprocessing

The interpretation of emotions based on brain activity is a complex and immature process, composed of the following stages: acquisition, preprocessing and classification. This section describe the process employed in this work to aquisition and preprocesisng EEG signal in order to prepare data to perform statistical classification experiments.

The signal acquisition was performed during the execution of a script task set aiming at evaluating the Facebook privacy features. At the end of the trial, the volunteer labeled the executed tasks as "easy" or "difficult" enabling EEG signals classification and process analysis. The acquisition was done using a dedicated system for biological signals acquisition from AdInstruments, with a 1kHz sampling rate, using eight electrodes for mapping brain activity in regions known as O, P, C and F related respectively to the visual cortex, spatial orientation, motor function and cognition.

The electrodes were connected through a specific cap, as show in Figure 2, and were filled with conductor gel to decrease the impedance with the scalp. The signal frequency was restricted to 50Hz to minimize the noise during acquisition. Based on the video process monitoring, the time spent on each task was define and applied as windowing to EEG data.

The activity segments were distributed into vectors with a maximum length of 10 s and the mean Power Spectral Density of each vector was extracted as feature into the following frequency ranges: PreAlpha (6 Hz to 8 Hz), Alfa (8 Hz to 12 Hz), Beta-1 (12 Hz to 14 Hz), Beta-2 (14 Hz to 18 Hz), Beta-3 (18 Hz to 25 Hz), Gamma-1 (25 Hz to 32 Hz) and Gamma-2 (32 Hz to 40 Hz), as suggested by Diez et al [13], resulting in 28 variables. In this study, these variables were called electrode-band, whose values were used to describe brain activity during the experiments.

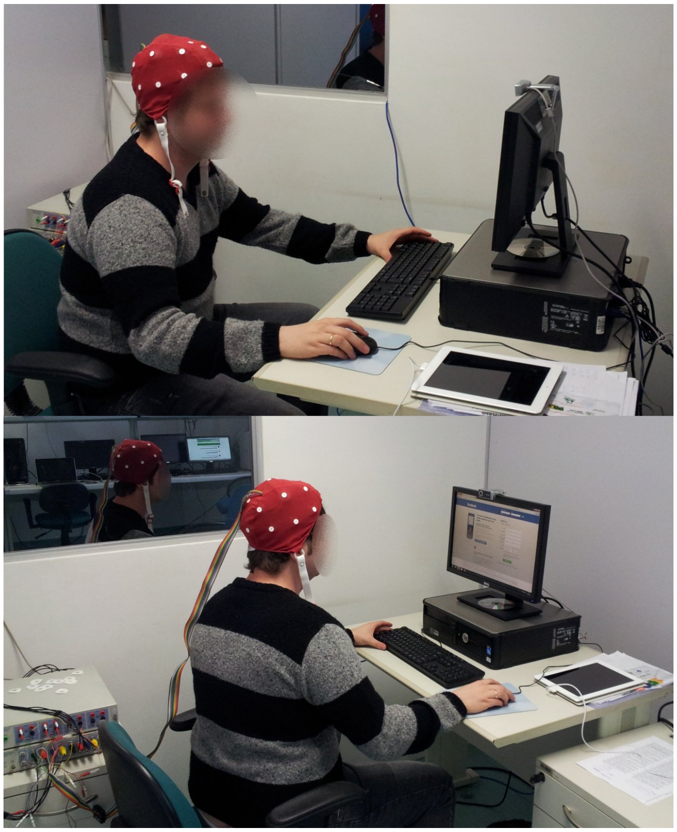

Fig. 2. EEG acquisition and usability test setup.

VI. Experiments and Results

In order to verify the benefits of the application of statistical techniques for usability tests, trials using the proposed methods were carried out. All experiments were conducted in Python 2.7, employing libraries Numpy, Scipy and Scikit-Learn.

A. *Hypothesis test: signal variability*

In this experiment, a hypothesis test was done to verify the statistical significance of changes in brain activity between the tasks defined by the user as “easy” or “difficult”. For each electrode-band, obtained in the pre-processing step, the Mann-Whitney U test was calculated, with a 95% confidence interval. Table I has highlighted cases where the null hypothesis was rejected (p-value $< .05$).

TABLE I
Mann-Whitney U test results for each electrode-band.

BANDS	**ELECTRODES**			
	P	**C**	**O**	**F**
PreAlpha	0,4579238	0,1056835	0,3112066	0,0064849
Alpha	0,2549813	0,0004875	0,3841640	0,0116818
Beta-1	0,0235355	0,0055816	0,3434209	0,0014289
Beta-2	0,0058000	0,0724194	0,3188574	0,0042024
Beta-3	0,0640959	0,0084775	0,3487170	0,0001915
Gamma-1	0,2626766	0,0272498	0,1772583	0,0000002
Gamma-2	0,0682482	0,0060259	0,2378651	0,3188574

B. *LDA and SVM classification*

The preprocessed data was classified twice with each selected method (LDA and SVM using the RBF Kernel to increse the dimensionality of the dataset). In the first setup, every electrode-band were used, and in the second one, only the thirteen that rejected the null hypothesis of the Mann-Whitney U test were used.

The Leave-One-Out method was chosen to validate the classifiers. In this method, training and test sets are generated from the dataset and each one is used to train the classifiers first and then test the classification result. Considering that the dataset has n total observations, each training set has $n-1$ observations and the test is made with the data that was taken from the dataset. In the end, it will be made n training and test process changing the datasets used in the classification.

The correct classification rate is shown in Table II, in which the values were obtained from counting how many correct classifications were made by each algorithm using the Leave-One-Out method described previously.

TABLE II
EEG signal classification results.

Electrode-band	**LDA**	**SVM**
28 (Original dataset)	62%	71%
13 (Preselected)	63%	65%

VII. Discussion

In the present work we performed a statistical inference test on the EEG signals obtained during a usability evaluation of Facebook, in which a volunteer performed tasks defined in a script and classified them as "easy"’ or "difficult".

In this context, the Mann-Whitney U test between samples from task groups labeled as “easy” or “difficult” was calculated for each electrode-band. For the electrode-band P-Beta1, P-Beta2, C-Alpha, C-Beta1, C-Beta3, C-Gamma1, C-Gamma2, F-PreAlpha, F-Alpha, F-Beta1, F-Beta2, F-Beta3 and F-Gamma1 the null hypothesis of equality between means was rejected. Therefore, in these cases, there were evidence of brain activity variation according to the difficulty of the tasks presented in the usability test script.

Finally, we compared the classification capacity of LDA and SVM on a Leave-One-Out test. Initially, we used the 28 original electrode-bands, and after only 13 electrode-bands whose null hypothesis was rejected at the Mann-Whitney U test. The classification rates showed a small advantage of the SVM method over LDA in the first experiment. However, LDA showed good discrimination capability of the samples with only 13 variables. Therefore, the electrode-bands P-Beta1, P-Beta2, C-Alpha, C-Beta1, C-Beta3, C-Gamma1, C-Gamma2, F-PreAlpha, F-Alpha, F-Beta1, F-Beta2, F-Beta3 e F-Gamma1 concentrated more information about the brain activity variation during the usability test conducted in this work, than other electrodes used in signal acquisition. It can be notice the predominance of the F and C electrode-bands related to cognition and motor functions, those extensively used to perform the proposed tasks and associated with the labeled emotions.

The cited works [2], [3] and [4], showed a statistical and spectral analysis comparing the differences between users surveys, user based tests and biological signals, but they did not use any classifier. Although, the similar studies [1] and [5] employed a lot of signal features to perform a classification experiments. Murugappan et al [1] reported a maximum average classification rate of 83.04% in the best experiment performed with K nearest neighbor (KNN) classifier using 62 channels. Athough, using only 24 channels and LDA with power spectral features, they obtained a maximum average classification rate about 57%. Koelstra et al [5] conducted a extensive statistical analisys between brain activity and participant's ratings. The results showed a significant correlation between both. However, the mean classification results obtained was about 50%. Therefore, comparing these classification rates with those attained by the present work, the results were interesting and relevant to emotion analysis context.

VIII. Conclusion

Despite the fact that this was a preliminary study, it showed the feasibility of using the EEG as a potential source of information to be added to software usability testing. However, it is necessary a larger number of samples allowing to confirm and generalize this results. In the future works we intend to apply another statistical analisys to not only classify the brain activity but to evaluate the intensity of emotional states from the users in a usability test. Thus we hope to provide more accurate resources to researchs in this area.

Additionally, we intend to use the results of EEG usability testing to provide relationship map of 48 emotions that people can feel. This is the EARL (Emotion Annotation and Representation Language) [14], which detected eight basic emotions, plus 8 advanced and 8 resulting in feelings. Mixing it all, human beings are capable of feeling 48 different emotions. The map will help organize the emotions of the user for User eXperience (UX) professionals.

References

[1] M. Murugappan, R. Nagarajan, and S. Yaacob, "Comparison of different wavelet features from eeg signals for classifying human emotions," *Journal of Medical and Biological Engineering*, vol. 31, no. 1, pp. 45–52, Maio 2010.

[2] M. Kimura, H. Uwano, M. Ohira, and K.-I. Matsumoto, "Toward constructing an electroencephalogram measurement method for usability evaluation," in *Proceedings of the 13th International Conference on Human-Computer Interaction. Part I: New Trends*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 95–104.

[3] H. Lee and S. Seo, "A comparison and analysis of usability methods for web evaluation: The relationship between typical usability test and bio-signals characteristics (eeg, ecg)," in *The proceedings of the 2010 DRS Montreal Conference*, D. Durling, Ed., 2010.

[4] H. Masaki, M. Ohira, H. Uwano, and K.-i. Matsumoto, "A quantitative evaluation on the software use experience with electroencephalogram," in *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*, ser. Lecture Notes in Computer Science, A. Marcus, Ed. Springer Berlin Heidelberg, 2011, vol. 6770, pp. 469–477.

[5] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Janeiro 2012.

[6] J. Rubin and D. Chisnell, *Handbook of Usability Testing: Howto Plan, Design, and Conduct Effective Tests*, 2nd ed. Wiley, May 2008. [Online]. Available: http://www.worldcat.org/isbn/0470185481

[7] P. T. Aquino Junior, "Picap: padrões e personas para expressão da diversidade de usuários no projeto de interação," Tese de Doutorado, Escola Politécnica, Universidade de São Paulo, São Paulo, SP, Brazil, 2008, date: 19 Apr. 2012. [Online]. Available: http://www.sheffield.ac.uk/eee/research/iel/research/face

[8] P. T. Aquino Junior and L. V. L. Filgueiras, "User modeling with personas," in *Proceedings of the 2005 Latin American conference on Human-computer interaction*, ser. CLIHC '05. New York, NY, USA: ACM, 2005, pp. 277–282. [Online]. Available: http://doi.acm.org/10.1145/1111360.1111388

[9] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Human Genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[10] S. Marsland, *Machine Learning: An Algorithmic Perspective*. NJ, USA: CRC Press, 2009.

[11] V. Vladimir and V. Vapnik, "The nature of statistical learning theory," 1995.

[12] TechSmith, "Morae," https://www.techsmith.com/morae.asp, 2012, date: 19 Apr. 2012.

[13] P. F. Diez, E. Laciar, V. Mut, E. Avila, and A. Torres, "Classification of mental tasks using different spectral estimation methods, biomedical engineering," in *Biomedical Engineering*, C. A. B. Mello, Ed. In-teh, 2009, vol. 1, pp. 287–306.

[14] M. Schröder, L. Devillers, K. Karpouzis, J.-C. Martin, C. Pelachaud, C. Peter, H. Pirker, B. Schuller, J. Tao, and I. Wilson, "What should a generic emotion markup language be able to represent?" in *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, ser. ACII '07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 440–451.