Transforming our World Through Design, Diversity and Education
G. Craddock et al. (Eds.)

doi:10.3233/978-1-61499-923-2-811

Method for Semi-Automated Evaluation of User Experience Using Brain Activity

Aleksander BAI[a,1] and Kristin S. FUGLERUD[a]
[a] *Norwegian Computing Center, Oslo, Norway*

Abstract. There is a large interest in user experience today, both from a usability and accessibility point of view. However, in order to verify what the users actually like and don't like, user testing must be conducted. Traditionally, user experience is measured retrospective with surveys and interviews, but this is not the most optimal approach since it does not measure user experience in the moment and it is prone for human error because of our inaccurate memory recollection. Here we propose a method that does semi-automated evaluation of user experience by utilizing electrophysiological equipment that monitors electrical activity of the brain. We describe an approach that together with brain activity monitoring will collect and quantify user experience in a non-intrusive manner. We demonstrate the method by showing how a low cost device can record brain activity during a user test, and auto-detect where the user has difficulties understand or navigating a solution. All this is done in an unsupervised manner, but an observer must still verify the feedback with the actual user to remove false positives. Our method is not limited to digital solutions and can also be used for evaluating user experience of physical installations.

Keywords. Usability, User experience, Universal design, Semi-automated, EEG

1. Introduction

First Traditionally user experience (UX) is measured through retrospective methods like surveys and interviews [1]. While these methods can give an in-depth understanding of the users' values, perceptions, and experiences, they are not very accurate when it comes analyzing details of the user experience. This is because humans are not very good at remembering details, even about events that have just happened. We are also very easy to manipulate and are highly influenced by others, and if questions are not asked precisely, the interviewer or survey might affect the answer from a participant [2].

Contextual interviews and think aloud are alternatives that can alleviate some of these weaknesses [3]. By performing the interviews in context, users will be able to recall more details. Think aloud is performed while the user is doing the task, but is not well suited for time sensitive tasks or when the user has difficulties in talking while doing. Think aloud is also something not everybody feel comfortable doing. So even though the motivation for asking a person about an experience is good, the outcome is not always very reliable or accurate.

There are several tools for automatic usability evaluation [4] and accessibility evaluation [5] based on guidelines. A challenge with these tools however, is that they do not necessarily have a good match with user experience evaluations with people [6]. Moreover, while there are many methods that uses an automated approach for data

[1] Corresponding Author, Aleksander Bai, Norwegian Computing Center, Oslo, Norway; E-mail: aleksander.bai@nr.no

collection from user experience [7], there are fewer examples of a well-defined method that combines automatic data collection and automatic evaluation of user experience with people. There is a call for more automated evaluation methods as a supplement for existing methods when measuring the effect of user experience [8].

Universal design (UD) is the design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design [9]. In UD, automatic data collection and evaluation can be particularly valuable for evaluating UX for participants that are not able to articulate themselves, for example due to speech impairments, intellectual impairments, or when being a foreigner that have not learned to speak the language in question fluently.

The remainder of the paper is organized as follows: After a summary of related work in Section 2 we describe our semi-automated method in Section 3. An evaluation of our method is shown in Section 4 before we discuss the possibility and implications in Section 5. We also summarize and highlight research directions in Section 7.

2. Related work

There are many studies that tries to use an automated approach for data collection for evaluation of user experience, quality of experience and usability. One of the most common approaches is to use eye-tracking in user experience design and evaluation to find gaze, heatmaps, area of interest and insight of where users are looking [10]. Other popular approaches for measuring user experience are pupil dilation as a method to measure cognitive load [11], skin conductance and heart-rate monitoring to measure stress [12] and facial emotion detection to detect emotions [13].

In affective computing there are some interesting research that use different methods to detect affection. In affective detecting the goal is to detect the user's emotions, for instance by using BCI and EEG [14], which is not exactly the same as user experience. The reason is that emotions must be translated into user experience, and the mapping is not trivial. For instance research has shown that people smile even though they answer questions incorrectly [15].

The use of brain activity and EEG monitoring has also been studied in relation to user experience [16]. However, in most cases it is required to train a machine learning algorithm in a supervised manner [17]. To our knowledge there are few automated and unsupervised methods that does not require labelled training data [18]. There are approaches that have fully automated methods for evaluating user experience [19], but they have not been tested with users and only on simulated data. Researchers also argue that data-driven methods, and triangulation of qualitative insights and quantifiable measures is important [20]. Thus a fully automated method is probably not feasible with the current technology.

3. Method

In order to do an automated evaluation of user experience, we need to actively monitor physiological and mental states of the users. Even though there are several studies that show that basic emotions can be detected from physiological sensors, the current available technology have trouble deducing subtle differences in emotions for a user, like the difference between slightly annoyed or mildly positive. These differences are

important from a user experience point of view, and part of the problem is that the physiological measurements looks similar. In the previous example, both emotions can results in a slightly elevated heart rate or increase of galvanic skin response [13]. In addition to being very individual from person to person, it is difficult to accurately distinguish the physiological data of different emotions from each other [21].

Studies have shown that when a person is angry, frustrated, excited and so on, she or he will in most cases show several physiological signs that can be monitored [22]. This is for instance how the lie detector works, by monitoring heart rate and galvanic skin response and comparing them against a baseline. However, some emotions are also harder to detect with physiological sensors, like concentration and fatigue. These states are easier to detect by monitoring the electrical activity of the brain, since they require much "brain power" [22].

We believe that a more promising approach, at least with the current technology, is to detect large differences in a person's cognitive load compared to a baseline. Our method utilize the research done within attention and fatigue brainwaves, and combines that with a mobile, low-cost EEG devices to deliver a semi-automated evaluated of user experience. In theory it could be fully automated, but so far the EEG technology is very sensitive to noise and individual differences in scalp and muscle configuration. The technology is able, however, to tell us that the participant experience a high level of concentration at a certain time, and it's up to the organizer to question the participant about why something was frustrating, demanding or exciting. Since we are using low-cost EEG devices that are mobile and allow the participant to move about, the ratio of signal to noise is quite high, and this will again result in several false positives. This is mainly because the EEG devices that is monitoring the brain activities will also pick up noise from scalp scratching from the dry electrodes, head movement, external noise like the power grid, muscle activities and such [23].

We have illustrated our concept in Figure 1, where our method uses EEG to measure brain activity to perform a semi-automated evaluation of user experience. More specific, the different brain waves are monitored constantly, and any significant increase in the brain waves will indicates that something has happened. This is done by first creating a baseline when the user is relaxed and not actively engaged in any activity. After a short baseline phase, the system will then be able to detect even small shifts in the brainwaves when the user is concentrating. The details behind our method is explained in Section 4.

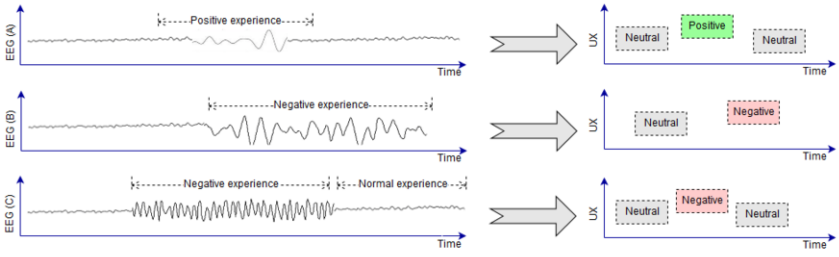

Figure 1. Method for semi-automated evaluation of user experience.

By using a mobile device for monitoring brain activity it is possible to use our method outside the lab and not only in front of a computer screen. Our method also makes it possible to measure user experience when testing and using physical installations.

However, when using EEG devices on users that are moving around, even more noise will be introduced, which support the need for a human organizer that can questioning the user and remove false positives.

4. Evaluation

To test if our method is feasible we have done a short evaluation of our method during a user test with eye-tracking. We tested a webpage for apartment rental. In this user test, we also asked the test users to wear a Muse headband [24]. The Muse device is small and portable, but have very few sensors. There are other consumer EEG headset available that are also portable, but Muse is one of the least intrusive devices [25]. In total 5 subjects were recruited and conducted a usability test with scenarios, where we used a Tobii T-60 eye tracking tool for detecting problems [26]. The participants were encouraged to think aloud.

At the end we performed an interview where we asked more about the specific issues they found and what elements they liked. Each user's session lasted between 25 and 40 minutes, and they were given five tasks to complete. Two of the tasks were very hard, and we expected them to experience trouble during those tasks, so we should be able to detect signs of high concentration.

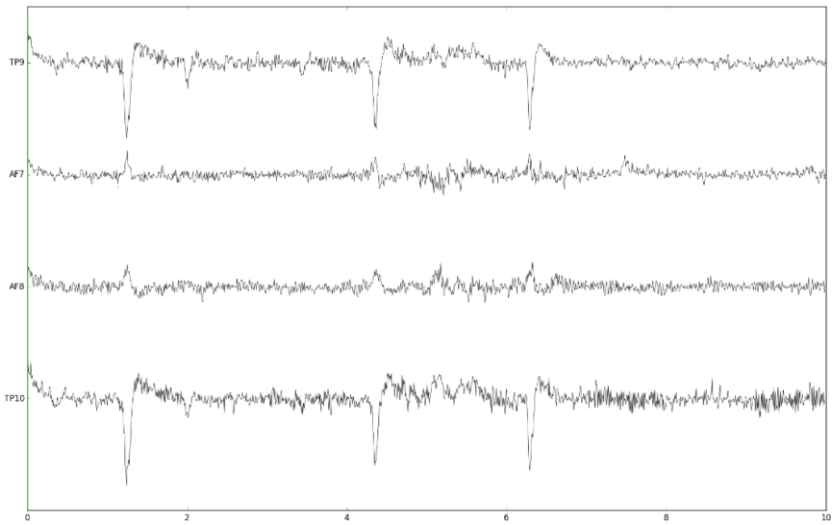

Figure 2. Raw EEG for a participant.

The raw EEG as shown in Figure 2 must be preprocessed before any high activity can be detected. The placement of sensors in Muse follows the international 10-20 standard [27], and have four sensors as indicated in Figure 2 (y-axis). The sensors AF7 and AF8 are located in the frontal part of the skull while TP9 and TP10 are located in front of each ear.

We also removed artifacts like blinking (shown in Figure 2 as downwards spikes) and muscle activity by using ICA [28]. We also applied a low-pass filter to remove inference from the power grid (50Hz in Europe), before we removed the baseline data

from the test data. Finally a spectrogram was produced, as shown in Figure 3. This illustrate the frequency intensity for a particular user in a session, and we can see that there are spikes where the brain activity is higher than the rest. It has been shown in multiple studies that activity in the different frequency bands is associated with high concentration [29], and this can also be seen in Figure 3 as vertical lines that almost run across the whole frequency range.

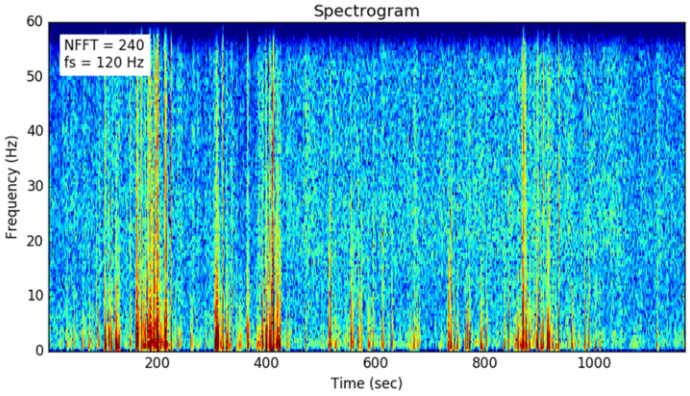

Figure 3. Spectrogram of a participant's EEG.

Based on the processed EEG data and a spectrogram, we applied chunking of the spectrogram. This means that we grouped larger sections together and averaged over them. We tried multiple intervals, and got good results with 10 and 30 seconds. After chunking was applied we did a unsupervised clustering with k-means and two classes [18]. We also tried with other classes, in order to get more subtle changes, but could not get any promising results with more than two classes. Figure 4 show the results of 10 seconds chunking and auto-detected difficulties. As the figure show, the clustering approach is able to detect groups of difficulties that have high brain activity. Our approach is also very fast, and can be performed in just a few seconds.

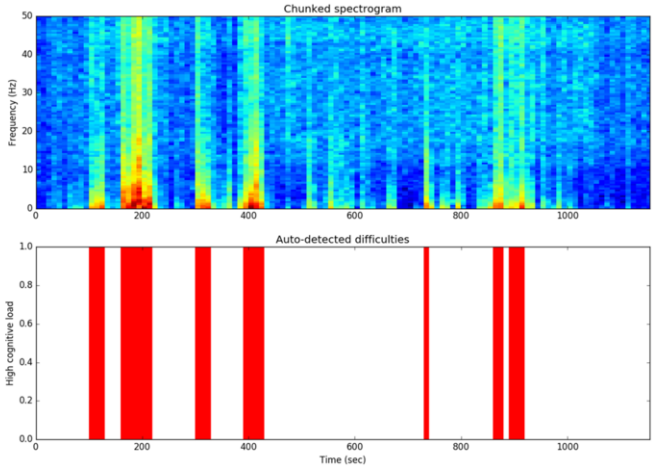

Figure 4. Spectrogram of an EEG with 10 seconds chunking and auto-detected difficulties.

In Figure 5 the same analysis with 30 seconds chunking is shown. In our experiment there was not any major benefits of using 30 seconds instead of 10.

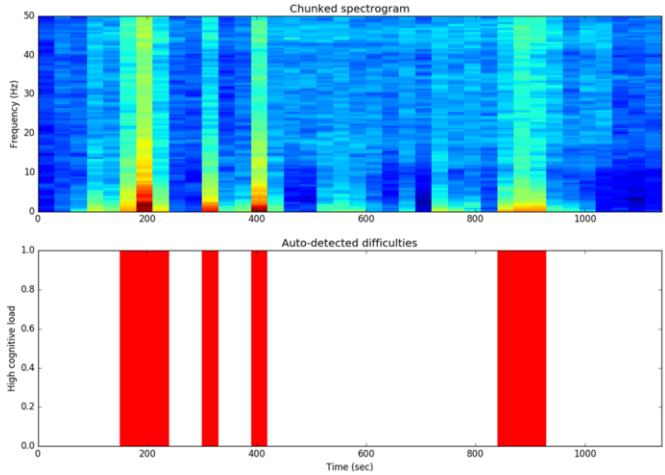

Figure 5. Spectrogram of an EEG with 30 seconds chunking and auto-detected difficulties.

A separate person went through the video recording of the subjects and tagged sections where the user seemed to have difficulties. For instance, when he or she was stating that "this was hard to understand" or "i am not sure how to proceed". We marked the start and end time where the subjects showed difficulties, and grouped together sections that were very close together. We refer to these tagged sections as observed difficulties. In Figure 6 the matching of auto-detected difficulties (10 seconds) are matched against observed difficulties.

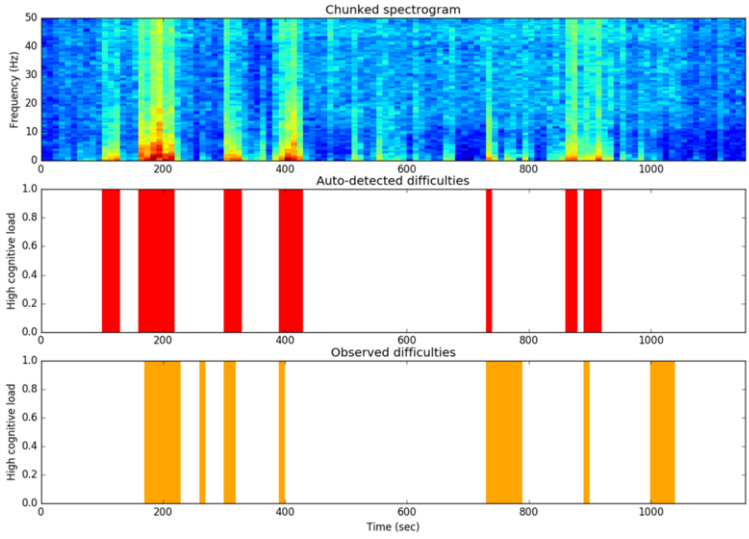

Figure 6. Matching auto-detected difficulties with observed difficulties.

For three of the subjects (subject 1 shown here) the match between auto-detected and observed difficulties were good with over 70\% accuracy of detecting observed difficulties. However, for two of the subjects the match was not accurate enough because of too much noise in the data. This was caused by bad connection of the Muse itself, which caused half of the EEG sensors to produce too much noise. This made our preprocess stage very difficult, and we were therefore not able to get any good matching. Two of the participants had quite a lot of hair which we suspect caused troubles, and the last had glasses that made the connection behind the ears unreliable. In retrospective we should have examined the data quality and connection better, both during setup and under the user testing.

5. Discussion

In our user testing we discovered that the registration process for the webpage was much harder for older people than the participant expressed. We did not have real-time analysis of the EEG during the user trial, but we performed the analysis afterwards and spotted very high activity during the registration process. We then went back and did analysis on the eye tracking data, and based on both pupil dilation and saccade patterns [30] it seemed that the participant had bigger problems with the registration process than the user trial feedback and interview indicated. We did not, however, manage to confirm this with the user since the user testing was completed.

Several studies [7] have also found that physiological measures like EEG are not yet widely accepted in the evaluation of the user experience and that people prefer qualitative approaches instead of quantitative. However, we support the idea that for data-driven methods, a triangulation of qualitative insights and quantifiable measures is important [31]. Our work fits well into that approach, but must be verified and extended to provide real-time analysis.

There are many applications that can benefit from a semi-automated evaluation of user experience. Webpages as used in our user testing is an obvious candidate, but it can also be extended to product evaluation. During our interviews we asked the users how they felt about wearing an EEG headband, and four out of five said it didn't bother them, and two users said they actually forgot that they were wearing an EEG headband. The last user said it caused some problems with the glasses and we probably should have used more time to configure the headband for that particular user.

Another benefit by using a mobile EEG monitoring device, is that our method can be expanded to evaluation of user experience in physical installations where the user moves around and interacts with a service or installation. It can also be an important method in contexts where the user is unable to verbalize or explain exactly what or when they experienced a problem. This could be in situations where time is critical, where the task requires high concentration, or the user has a cognitive or speech impairment. The method is also a good candidate for evaluations in the Internet of Things [32] or ambient intelligent environments [33].

There are several aspect that we would like to address better in the future. First and foremost should similar devices with better signal-to-noise ratio like the EPOC+ [34] or OpenBCI [35] should be explored. Since they have more sensors, they should produce better results (better signal to noise ratio) while still being easy to setup and operate. In addition, better feature extraction algorithms should be explored.

Another interesting research path is to figure out how to accurate detect positive and negative emotions. There are studies that have shown that it's possible to detect emotions [21]. A method which tries to combine a portable EEG device with emotion detection would be interesting. Detection of positive emotions can make it possible to identify positive user experiences and enhance those.

Finally we would like to see our approach verified with physical installations. Measuring user experience is a complex and resource demanding task with the current methodologies, and in particular for physical buildings where the users might move around for a longer period of time. With regard to universal design and evaluation of physical buildings, we believe that physiological monitoring in combination with EEG is a very promising path that must be studied more.

6. Limitation

Our study of was limited in number of participants and the size of the evaluated applications. Hence, future work should verify the method for more users and in different settings.

We did not ask the users to confirm our findings and verify the automatic detection of difficult segments. Ideally we would have the users themselves provide the ground truths by watching the video shortly after the session was over.

7. Conclusion

In this study we have proposed a semi-automated method for evaluation of user experience. Our method uses EEG equipment to auto-detect high brain activity, which indicates high cognitive load and difficulties. These difficulties are found unsupervised, but must be verified by a human to remove false positives. We have verified our approach with a low-cost device called Muse, and because of the signal to noise ration a short baseline period is beneficial before the actual user testing starts. It is possible to use our approach without a baseline, but more noise reduction techniques will then be required.

We also group time windows together to create what we call chunks, since this makes our unsupervised method more robust. However, we need more studies with more participants and different equipment to evaluate how reliable and durable our approach is. Finally we would also like to see our method extended to user testing of physical products and installations.

References

[1] A. P. O. S. Vermeeren, E. L.-C. Law, V. Roto, M. Obrist, J. Hoonhout, and K. Väänänen-Vainio-Mattila, "User experience evaluation methods: current state and development needs," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 2010, pp. 521–530.

[2] K. Erdmann, R. Volbert, and C. Böhm, "Children report suggested events even when interviewed in a non-suggestive manner: what are its implications for credibility assessment?," *Appl. Cogn. Psychol.*, vol. 18, no. 5, pp. 589–611, 2004.

[3] R. Hartson and P. S. Pyla, *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier, 2012.

[4] A. Dingli and J. Mifsud, "USEFul: A Framework to Mainstream Web Site Usability Through Automated Evaluation," *Int. J. Hum. Comput. Interact.*, vol. 2, no. 1, p. 10, 2011.
[5] A. Aizpurua, S. Harper, and M. Vigo, "Exploring the relationship between web accessibility and user experience," *Int. J. Hum. Comput. Stud.*, vol. 91, pp. 13–23, 2016.
[6] M. Vigo, J. Brown, and V. Conway, "Benchmarking web accessibility evaluation tools: measuring the harm of sole reliance on automated tests," *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. ACM, Rio de Janeiro, Brazil, pp. 1–10, 2013.
[7] C. L. B. Maia and E. S. Furtado, "A systematic review about user experience evaluation," in *International Conference of Design, User Experience, and Usability*, 2016, pp. 445–455.
[8] A. Bai, H. C. Mork, T. Halbach, K. S. Fuglerud, W. Leister, and T. Schulz, "A Review of Universal Design in Ambient Intelligence Environments."
[9] R. Mace, "What is universal design," *Cent. Univers. Des. North Carolina State Univ. Retrieved Retrieved Novemb.*, vol. 19, p. 2004, 1997.
[10] J. R. Bergstrom and A. Schall, *Eye tracking in user experience design*. Elsevier, 2014.
[11] D. Wendt, T. Koelewijn, A. A. Zekveld, and T. Lunner, "Investigating the effect of competing talkers on speech processing load as shown by task evoked pupil dilation," in *the 3rd international Conference on Cognitive Hearing Science for Communication (CHSCOM), Linköping, Sweden, June 2015 14-17, 2015*, 2015.
[12] A. Liapis, C. Katsanos, D. Sotiropoulos, M. Xenos, and N. Karousos, "Recognizing emotions in human computer interaction: studying stress using skin conductance," in *Human-Computer Interaction*, 2015, pp. 255–262.
[13] W. Albert and T. Tullis, *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes, 2013.
[14] C. Mühl, D. Heylen, and A. Nijholt, "Affective brain-computer interfaces: neuroscientific approaches to affect detection," in *Oxford Handbook of Affective Computing*, Oxford University Press Oxford, 2015, pp. 217–232.
[15] W. Leister, I. Tjøstheim, T. Schulz, G. Joryd, A. Larssen, and M. de Brisis, "Assessing visitor engagement in science centres and museums," *Studies*, vol. 17, no. 18, p. 15, 2016.
[16] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 186–197, 2010.
[17] A.-N. Moldovan, I. Ghergulescu, S. Weibelzahl, and C. H. Muntean, "User-centered EEG-based multimedia quality assessment," in *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*, 2013, pp. 1–8.
[18] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised learning," in *The elements of statistical learning*, Springer, 2009, pp. 485–585.
[19] A. R. Marathe, J. R. McDaniel, S. M. Gordon, and K. McDowell, "Confidence-Based State Estimation: A Novel Tool for Test and Evaluation of Human-Systems," in *Advances in Human Factors in Robots and Unmanned Systems*, Springer, 2017, pp. 291–303.
[20] I. Pettersson, F. Lachner, A.-K. Frison, A. Riener, and A. Butz, "A Bermuda Triangle?-A Review of Method Application and Triangulation in User Experience Evaluation," 2018.
[21] L. E. Nacke, "Games user research and physiological game evaluation," in *Game user experience evaluation*, Springer, 2015, pp. 63–86.
[22] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven, "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviat. Space. Environ. Med.*, vol. 78, no. 5, pp. B231--B244, 2007.
[23] M. Abujelala, C. Abellanoza, A. Sharma, and F. Makedon, "Brain-ee: Brain enjoyment evaluation using commercial eeg headband," in *Proceedings of the 9th acm international conference on pervasive technologies related to assistive environments*, 2016, p. 33.
[24] Muse, "Muse." http://www.choosemuse.com/.
[25] L. Galway, P. McCullagh, G. Lightbody, C. Brennan, and D. Trainor, "The potential of the brain-computer interface for learning: a technology review," in *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on*, 2015, pp. 1554–1559.
[26] Tobii AB, "Tobii T60." https://www.tobiipro.com/product-listing/tobii-t60-and-t120/.
[27] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10--20 system electrode placement," *Electroencephalogr. Clin. Neurophysiol.*, vol. 66, no. 4, pp. 376–382, 1987.
[28] R. Vigário, J. Sarela, V. Jousmiki, M. Hamalainen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 5, pp. 589–593, 2000.

[29] N.-H. Liu, C.-Y. Chiang, and H.-C. Chu, "Recognizing the degree of human attention using EEG signals from mobile sensors," *Sensors*, vol. 13, no. 8, pp. 10273–10286, 2013.
[30] J. N. Sari, R. Ferdiana, P. I. Santosa, and L. E. Nugroho, "An eye tracking study: exploration customer behavior on web design," in *Proceedings of the International HCI and UX Conference in Indonesia*, 2015, pp. 69–72.
[31] K. S. Fuglerud, "Inclusive design of ICT: The challenge of diversity," University of Oslo, Faculty of Humanitites, 2014.
[32] F. Xia, L. T. Yang, L. Wang, and A. Vinel, "Internet of things," *Int. J. Commun. Syst.*, vol. 25, no. 9, pp. 1101–1102, 2012.
[33] J. Alves Lino, B. Salem, and M. Rauterberg, "Responsive environments: User experiences for ambient intelligence," *J. Ambient Intell. Smart Environ.*, vol. 2, no. 4, pp. 347–367, 2010.
[34] Emotiv, "Emotiv EPOC+." https://www.emotiv.com/epoc/.
[35] OpenBCI, "OpenBCI." https://www.openbci.com/.