Detecting and Comparing Brain Activity in Short Program Comprehension Using EEG

Martin K.-C. Yeh
College of Information Sciences and Technology
Penn State University, Brandywine
martin.yeh@psu.edu

Dan Gopstein
Department of Computer Science and Engineering
New York University
dgopstein@nyu.edu

Yu Yan
College of Education
Penn State University, University Park
yanyu@psu.edu

Yanyan Zhuang
Department of Computer Science
University of Colorado, Colorado Sprints
yzhuang@uccs.edu

***Abstract*—Program comprehension is a common task in software development. Programmers perform program comprehension at different stages of the software development life cycle. Detecting when a programmer experiences problems or confusion can be difficult. Self-reported data may be useful, but not reliable. More importantly, it is hard to use the self-reported feedback in real time.**

In this study, we use an inexpensive, non-invasive EEG device to record 8 subjects' brain activity in short program comprehension. Subjects were presented either confusing or non-confusing C/C++ code snippets. Paired sample t-tests are used to compare the average magnitude in alpha and theta frequency bands. The results show that the differences in the average magnitude in both bands are significant comparing confusing and non-confusing questions. We then use ANOVA to detect whether such difference also presented in the same type of questions. We found that there is no significant difference across questions of the same difficulty level. Our outcome, however, shows alpha and theta band powers both increased when subjects are under the heavy cognitive workload. Other research studies reported a negative correlation between (upper) alpha and theta band powers.

Keywords—computer programming; electroencephalograph; EEG

I. INTRODUCTION

Software design includes complex cognitive tasks including program comprehension where symbols and expressions are to be translated and combined to create the expected outcome. Program comprehension is performed at different stages of the software development life cycle and at different times. It is essential for software developers to perform program comprehension to create software and to avoid flaws. This study is to understand whether programmers react differently to short C/C++ code snippets of different types through recording and analyzing their brain activity and whether the brain activity measure is consistent with the type of code snippet (confusing vs. non-confusing).

To test our hypothesis that brain waves are different when people are solving code snippets, we created two versions of code snippet, one is confusing, hence more difficult to come up with an answer, and the other is non-confusing, hence easier to solve, based on six features of C/C++. The pair of code snippets in each feature are essentially equivalent. Subjects were asked to solve six pairs, twelve in total, of code snippets. These questions have been tested by programmers to confirm that the confusing code snippets are indeed confusing—subjects showing significantly lower accuracy and longer time on task [1].

In addition to the code snippets, we asked subjects to indicate how difficult the question they just saw was and how confident they were about the answer they entered. The self-reported data can provide data to understand how subjects perceive each code snippet.

To record subjects' brain activity, we used an inexpensive, non-invasive, consumer-grade EEG (electroencephalograph) device manufactured by Emotiv called Epoc+. The total cost of the device and software is less than one thousand dollars.

It is difficult to capture the moment when a programmer experiences problems or confusion. These type of data are typically self-reported. Alternatively, the difficulty of the code snippets can be assessed by scoring the outcome, either by accuracy or quality. Either method, however, fails to provide just-in-time feedback for further applications. Moreover, a code snippet may be confusing to one person but not confusing to another. Although it is possible to test different features by using a large number of human subjects, EEG signals provide a way to detect whether a code snippet is confusing or not.

As non-invasive EEG devices becoming more accessible and signal processing techniques becoming more advanced, it is now possible to collect physiological data that reflects cognitive workload during learning and problem-solving processes. This can be particularly useful for educational applications such as intelligent tutoring systems.

II. RELATED WORK

The EEG signal reflects an electrical current in the brain that can be recorded using invasive (electrodes placed cortical surface) and non-invasive (electrodes placed on the scalp).

This project is supported by the National Science Foundation under Grant No. 1444827.

978-1-5090-5920-1/17/$31.00 ©2017 IEEE

Different devices provide different spatial densities (number of electrodes) and resolutions (sampling rate). Interested readers can read [2]–[4] for more details and background knowledge about EEG. We select studies that are closely related to this paper and discuss them below.

A. Brain Waves as Indicators

1) Theta Frequency

The theta frequency band (4 – 8 Hz) is often associated with the degree of mental process, cognitive workload, or working memory load. In a study, Raghavachari et al. [5] aimed to determine the relation between working memory load and the power of EEG signal in the theta frequency band. They recorded four subjects' EEG signals while the subjects performed the Sternberg task, which is a non-spatial task, using iEEG devices (an invasive method that places a small array of electrodes on the cortical surface.) They found that the amplitude of theta frequency band increased at the beginning of the trial and remain strong throughout the trials. Another earlier study [6] also reported that an increase in theta band power was related to working memory load. Both studies suggest that theta frequency power is positively related to the working memory workload for non-spatial tasks. The task we used in the study is also non-spatial (program comprehension.) However, we are aiming to discover whether the non-invasive EEG that covers a larger area of the brain than iEEG does can produce similar outcomes because signals from non-invasive methods contain more noise and interference (e.g., eye blinks, muscle movements, signals travel from neurons to the skull.)

2) Alpha Frequency

Alpha frequency band (8 – 13 Hz) is one of the earliest frequency bands studied for making connection between EEG signals and brain activities. Similar to theta band power, alpha band power also changes in relation to working memory load and task performance. However, theta and alpha band powers interact with working memory load in an opposite way, i.e., when alpha band power increases, theta band power decreases [7]. In addition, researchers have found that the range of alpha frequencies differ by individual due to a wide range of factors such as age [7], memory performance [8], head size [9], etc. Normally, the alpha frequency band is analyzed in sub-bands (two Hz in each band): lower 1 alpha, lower 2 alpha, and upper alpha. Among them, upper alpha is the one that has been discussed the most and used for EEG analysis related to cognitive performance. Upper alpha band normally is defined as the frequency range from the *individual alpha frequency (IAF)* to *IAF + 2* Hz. In our study, we used broad alpha frequency band (8 – 13 Hz) instead of the upper alpha band because we do not have subjects' ages to calculate their IAFs.

3) Event-Related Desynchronization/Synchronization

EEG signals are inherently noisy and hard to analyze. One method called Event-Related Desynchronization (ERD) is often used in areas related to cognitive workload [10], [11]. ERD shows a time period that neurotic oscillation does not synchronize, which causes the amplitude to be weaker than when neurons oscillate synchronically. On the other hand, Event-Related Synchronization (ERS) is similar to ERD except that ERS is when neurons exhibit synchronized oscillation, which increases the strength of amplitude.

To calculate ERD, the amplitude during an event is compared with the amplitude from a wakeful, restful state. ERD is essentially the change of power in percentage from the restful state to the time when the stimulus is presented. The formula of ERD can be found in [12]. ERD/ERS is mentioned briefly here because of its popularity and for discussing related work. Our work, however, does not use this analysis because we do not have a wakefulness state as a reference for calculating ERD.

B. Applications of EEG

Typically, two methods can be used to assess people's cognitive effort. A traditional way is asking questions in surveys, which depends on people's subjective justification [13]. NASA Task Load Index (NASA-TLX) is an example instrument used in this method. Another method is using physiological measures, such as EEG devices, to directly assess cognitive load and awareness [14]. Many studies have used EEG devices to measure learner's cognitive load while learning information or solving problems, and the evidence showed that using EEG devices has some merits. For example, Antonenko and Niederhauser [15] used EEG data (alpha, beta, and theta bands) to determine the effect of hypertext leads on subjects' cognitive load and learning. They also measured cognitive load by collecting subjective data using a mental effort scale. The result indicated that using hypertext lead to lower cognitive load and resulted in better learning outcomes than links without leads. However, these differences only showed up when using alpha, beta, and theta measures in EEG data. There were no significant differences in the subjective measures. Antonenko and Niederhauser argued that the self-reported mental effort measure reflected the overall load and was associated closely with one specific type of load (e.g. intrinsic load) while EEG data was sensitive and could catch the change in instantaneous load and germane load.

An earlier study conducted by Gere and Jauscvec [16] investigated the differences in cognitive processes when subjects were learning information presented in different formats (text or multimedia) by using EEG data. The alpha power amplitude was calculated to measure the level of brain activity. They reported that text presentations showed higher cognitive load over frontal lobes (verbal processing), while video and pictures presentation displayed higher brain activity in occipital and temporal areas (visualization processing). They also reported that gifted students showed less mental activity.

Recently, EEG data have been used with tutoring/learning system to improve subjects learning performance. For example, Beal and Galan [17] used EEG to measure students' attention and cognitive workload while solving math problems in a tutoring system. They reported that students' performance (failure or success) could be correctly predicted by using EEG data, and EEG data also correlated with students' self-report of problem difficulty. Similarly, Chen and Huang [18] developed an attention-based self-regulated learning system using EEG devices. Sustained attention values were generated based on the real-time EEG data were recorded and then sent to the learning system. They reported a strong positive correlation between sustained attention and reading comprehension performance.

Researchers also used EEG devices to investigate different levels of expertise in programming. Crk, Kluthe and Stefik [12] used the EEG from when programmers were solving Java code snippets. ERD was calculated in alpha and theta bands as a measure of cognitive demands. Their results showed that EEG data can differentiate programmers with different level of expertise.

C. Confusing Code

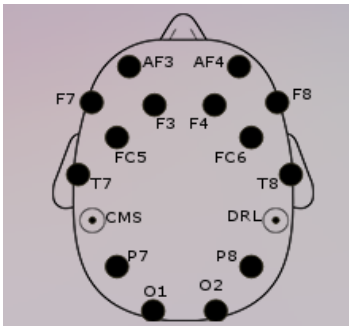
One of the oldest topics in software engineering is code comprehension. Recent work has moved towards building empirical and objective models of this comprehension. In particular, the Atoms of Confusion project has identified tiny pieces of code that have the ability to confuse programmers [1]. Candidates for these atoms of confusion were extracted from known confusing code, winners of the International Obfuscated C Code Contest. They were selected specifically to be as small as possible, but still exhibited confusion. A human-subjects experiment with 73 participants validated the ability of those tiny code snippets to confuse programmers. Subjects were shown pairs of minimal code snippets, on average only 6 lines for a complete program. Of these pairs, both programs would perform the same computation, but used different code to accomplish the task. One of the snippets in each pair was obfuscated, taken from the IOCCC winner, we refer to this type of snippet as "confusing". The other snippet was simplified to produce the same output without using the confusing construct, we refer to this type of snippet as "non-confusing". Programmers were asked to evaluate each code snippet by hand and record the output of each program. The results of this experiment showed that many of the atom candidates caused programmers to make errors at rates significantly higher than the simplified code. The data from that project indicated several very small patterns in code that dramatically increase a programmer's likelihood of misunderstanding a piece of code.

III. Instruments and Procedure

In our study, the subjects are eight undergraduate or graduate students who had taken at least one semester of C/C++ coursework (self-reported). After the experiment was explained to the subjects and consent form was signed, the first step was to fit the EEG device on the subject's head. Then, the subject used a web-based application that we created using jsPsych [19] to record their answers and the timestamp when each code snippet was shown to the subject. We customized it and created plugins to meet our needs such as syntax highlighting and sliders to report answer confidence and difficulty. jsPsych has timing data for us to calculate the duration when the subject was exposed to each page, which was used to find out which stimulus the subject was looking at.

The application first showed an instruction page, then a sample question so that the subject could practice how to use the interface. Once the subject completed the practice and had no further questions, he/she was shown one code snippet, followed by one self-report on the difficulty of the question and then the confidence of his/her answer. This cycle of one code snippet followed by two self-report questions repeated until all twelve code snippets (mixed order of six confusing and six non-confusing counterparts) were answered.

Fig. 1. Electrode position of Emotiv Epoc+ device when the neuroheadset is not turned on. (When the neuroheadset is fitted and connected with the TestBench, the strength of each electrode is indicated by a color, green representing a good connection.)



During the experiment, the experimenter used another laptop to run TestBench, an EEG application from the vendor, to record the subject's EEG signals wirelessly. TestBench can output edf (European Data Format) and CSV (Common Separate Value). It also shows the strength of each channel in real time. EPoc+ has 14 channels (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) (Fig. 1.) with 128 Hz or 256 Hz sampling rate.

IV. Data Analysis

We imported the edf files into the R statistical analysis package. The analysis was done using signals from 8 channels that are related to cognitive load: AF3, AF4, F3, F4, F7, F8, FC5, and FC6. Signals were processed by first using a band pass filter between 0.16 and 13 Hz. The lower frequency is recommended by the EEG vendor to remove DC offset. The higher frequency of the band pass filter is because 13 Hz was the highest frequency we used. We then marked all amplitudes that were either greater than 200 μv or less than -200 μv as NA because signals outside of this range represent high noise [12].

To see whether there is a significant difference in terms of neuron synchronization during program comprehension, we used Fourier transform to convert the signal to the frequency domain. After using FFT, we separated the signal by question and into two groups: confusing and non-confusing. Signals that fell outside of the target time period were not included in the analysis. Means of magnitude were calculated for each question and for both confusing questions and non-confusing questions as a group on selected channels.

V. Results

A. Comparing magnitude in alpha and theta band between confusing questions and non-confusing questions

Paired sample t-tests (two tailed) were used to determine whether there is a significant difference in EEG magnitude between confusing questions and non-confusing questions. The means, standard deviations, and t-tests statistics are shown in Table I (alpha band) and Table II (theta band). Since multiple t-

tests were performed for each channel, a Bonferroni correction was used to determine the significance level to control for the inflation of Type I error. The alpha level was set to be .006 (α = .05/8) for each individual test. As can be inferred from Table I and Table II, confusing questions were associated with significant higher alpha and theta magnitude on most of the channels (p<.006). The alpha magnitude of confusing questions were 1.6 to 2.3 times as high as those of non-confusing questions. Similarly, the theta magnitude of confusing questions were 1.6 to 2.1 times as high as those of non-confusing questions. The magnitude differences in channel FC5 and FC6 were the largest (2 to 2.3 times) among all eight channels, both in alpha and theta band.

TABLE I. MEANS, STANDARD DEVIATIONS, AND PAIRED SAMPLE T-TEST (DF=7) IN ALPHA BAND MAGNITUDE.

	Confusing questions		Non-confusing questions		t-test	
Channel	***M***	***SD***	***M***	***SD***	***t***	***p***
AF3	304108.9	231830.6	190650.6	174916.0	3.08	0.018
AF4	291101.6	189488.3	173006.8	145355.4	4.71	0.002
F3	130961.4	89497.9	67764.0	52015.6	4.10	0.005
F4	146566.7	91491.4	89355.2	72142.0	4.46	0.003
F7	280277.6	383406.7	173060.1	265694.2	2.51	0.041
F8	397653.6	470870.7	246638.7	330333.7	2.96	0.021
FC5	119251.6	61383.2	51189.6	33183.3	4.42	0.003
FC6	198822.7	109836.6	92864.5	71200.5	4.32	0.004

TABLE II. MEANS, STANDARD DEVIATIONS, AND PAIRED SAMPLE T-TEST (DF=7) IN THETA BAND MAGNITUDE.

	Confusing questions		Non-confusing questions		t-test	
Channel	***M***	***SD***	***M***	***SD***	***t***	***p***
AF3	2583896.0	2656077.0	1536269.0	1779286.0	2.92	0.022
AF4	2547066.0	2306233.0	1411309.0	1617149.0	4.13	0.004
F3	797148.2	522820.2	394700.5	262533.5	3.52	0.010
F4	822321.8	479793.7	470026.1	319352.7	3.18	0.016
F7	2167013.0	3088490.0	1297680.0	2139929.0	2.44	0.045
F8	2591067.0	3327303.0	1575802.0	2431971.0	3.05	0.019
FC5	815413.1	549534.7	381596.2	327352.2	3.73	0.007
FC6	1146348.0	744481.7	559359.1	409597.3	4.50	0.003

B. Comparing magnitude in alpha and theta band within confusing questions and non-confusing questions

In the previous section (Section V.A.), we reported that there were significant differences in subjects' brainwaves when they were solving confusing or non-confusing questions. To investigate whether this effect is caused by the questions within the group instead of by the question type, we performed the following ANOVA tests.

Several one-way ANOVA with repeated measures were conducted to determine differences in alpha and theta magnitude when subjects were solving the different questions in the same confusing group. The between-subject factor is the different questions in the same confusing group. The Greenhouse-Geisser correction was used to account for any violation of the sphericity assumption.

We found no significant differences in subjects' alpha or theta magnitude when they were solving the six confusing questions or six non-confusing questions. The results were consistent across all eight channels. This indicates that subjects would have similar alpha and theta magnitude when solving programming questions with similar confusing level (difficulty level). It also validates the findings from previous analysis (Section V.A), that the differences found in the average alpha and theta magnitude between confusing and non-confusing questions are associated with the difficulty of the questions.

C. Absolute power and subjects' performance

Previous studies suggest that a large reference band power is associated with a large amount of desynchronization (alpha suppression) during task performance. Klimesch [7] pointed out that subjects with a good memory showed significantly stronger power in the upper alpha band.

A Pearson correlation was calculated to determine if the absolute power in the broad alpha band could predict subjects' performance. The subjects' performance was measured by the total number of correct answers. The correlation between subjects' performance and broad alpha power is r=0.72 ($p<0.05$). The correlations remain the same when calculated with the alpha power when solving confusing questions (r=0.70), or with alpha power when solving the non-confusing questions (r=0.73, $p<0.05$).

VI. CONCLUSION

In this work, we use an inexpensive, non-invasive EEG device to record subjects' brain activity during program comprehension and analyze the signals in the frequency domain. Overall the outcome is encouraging and has the potential for educational applications. Firstly, our analysis shows in both broad alpha and theta bands, the average band power (magnitude) are larger when solving confusing code snippets than when solving non-confusing code snippets. This indicates either more neurons are active or neurons oscillate in harmony. Moreover, there is no statistical difference among solving the same type of code snippet in the average magnitudes. This indicates that the magnitude is positively correlated to cognitive workload. Our work demonstrates that alpha and theta band powers can be used to differentiate the type of code by simply recording EEG signals on the scalp. Intelligent tutoring systems can use EEG as an input to provide detailed explanations, extra practices, additional examples, or select different instructional strategies.

Secondly, the results also exhibit that broad alpha band powers can be used to gauge subject's performance. This data can provide another modality for identifying experts or experienced users.

VII. FUTURE WORK

There are several areas we wish to improve in our future study. First, we did not add a long enough break between each question. Neuron oscillation is time sensitive and takes time to reflect the effect induced/evoked by the stimulus, therefore, adding a longer break between questions can potentially increase accuracy. Second, we did not collect subject age, which costs us the opportunity to calculate the peak alpha frequency [20] and calculate the upper alpha band for analysis because the peak alpha frequency is calculated based on age.

ACKNOWLEDGMENT

We would like to thank Justin Cappos, Chris Dancy, Korey MacDougall, and Frank Ritter for helping us improve the study. We also want to thank Asad Azemi and Tim Niller for advising us on signal processing.

References

[1] D. Gopstein *et al.*, "Understanding Misunderstandings in Source Code," in *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 2017.

[2] R. Stern, W. Ray, and K. Quigley, "Brain: Electroencephalography and imaging," in *Psychophysiological recording*, 2001, pp. 79–105.

[3] F. Lopes da Silva, "EEG: Origin and Measurement," in *EEG-fMRI: Physiological Basis, Technique and Applications*, C. Mulert and L. Lemieux, Eds. Berlin: Springer-Verlag, 2010.

[4] D. Millet, "The Origins of EEG," in *7th Annual Meeting of the International Society for the History of the Neurosciences (ISHN)*, 2002.

[5] S. Raghavachari *et al.*, "Gating of human theta oscillations by a working memory task," *J. Neurosci.*, vol. 21, no. 9, pp. 3175–3183, 2001.

[6] C. Tesche and J. Karhu, "Theta oscillations index human hippocampal activation during a working memory task," *Proc. Natl. Acad. Sci.*, vol. 97, no. 2, pp. 919–924, 2000.

[7] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain Res. Rev.*, vol. 29, no. 2, pp. 169–195, 1999.

[8] W. Klimesch, "EEG-alpha rhythms and memory processes," *Int. J. Psychophysiol.*, vol. 26, no. 1, pp. 319–340, 1997.

[9] P. L. Nunez, L. Reid, and R. G. Bickford, "The relationship of head size to alpha frequency with implications to a brain wave model," *Electroencephalogr. Clin. Neurophysiol.*, vol. 44, no. 3, pp. 344–352, 1978.

[10] G. Pfurtscheller and F. Lopes da Silva, "Event-Related EEG/MEG Synchronization and Desynchronization: Basic Principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, 1999.

[11] G. Pfurtscheller, "Graphical display and statistical evaluation of event-related desynchronization (ERD)," *Electroencephalogr. Clin. Neurophysiol.*, vol. 43, no. 5, pp. 757–760, 1977.

[12] I. Crk, T. Kluthe, and A. Stefik, "Understanding programming expertise: an empirical study of phasic brain wave changes," *ACM Trans. Comput.-Hum. Interact. TOCHI*, vol. 23, no. 1, p. 2, 2016.

[13] J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory*. New York, NY: Springer New York, 2011.

[14] P. D. Antonenko, F. Paas, R. Grabner, and T. van Gog, "Using Electroencephalography to Measure Cognitive Load," *Educ. Psychol. Rev.*, vol. 22, no. 4, pp. 425–438, Dec. 2010.

[15] P. D. Antonenko and D. S. Niederhauser, "The influence of leads on cognitive load and learning in a hypertext environment," *Comput. Hum. Behav.*, vol. 26, no. 2, pp. 140–150, Mar. 2010.

[16] I. Gerě and N. Jaušcvec, "Multimedia: Differences in cognitive processes observed with EEG," *Educ. Technol. Res. Dev.*, vol. 47, no. 3, pp. 5–14, 1999.

[17] F. C. Galán and C. R. Beal, "EEG estimates of engagement and cognitive workload predict math problem solving outcomes," in *User Modeling, Adaptation, and Personalization*, Springer, 2012, pp. 51–62.

[18] C.-M. Chen and S.-H. Huang, "Web-based reading annotation system with an attention-based self-regulated learning mechanism for promoting reading performance: Attention-based self-regulated learning mechanism," *Br. J. Educ. Technol.*, vol. 45, no. 5, pp. 959–980, Sep. 2014.

[19] J. R. De Leeuw, "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," *Behav. Res. Methods*, vol. 47, no. 1, pp. 1–12, 2015.

[20] W. Klimesch, "Memory processes, brain oscillations and EEG synchronization," *Int. J. Psychophysiol.*, vol. 24, no. 1, pp. 61–100, 1996.