2011 19th IEEE International Conference on Program Comprehension

Subvocalization – Toward Hearing the Inner Thoughts of Developers

Chris Parnin
College of Computing
Georgia Institute of Technology
Atlanta, Georgia USA
chris.parnin@gatech.edu

***Abstract*—Some of the most fascinating feats of cognition are never witnessed or heard by others, yet they occur daily in the minds of software developers practicing their craft. Researchers have desperately tried to glimpse inside, but with limited tools, the view into a developer's internal mental processes has been dim. One available tool, so far overlooked but widely used, has demonstrated the ability to measure the physiological correlates of cognition. When people perform complex tasks, sub-vocal utterances (electrical signals sent to the tongue, lips, or vocal cords) can be detected. This phenomenon has long intrigued researchers, some likening sub-vocal signals to the conduits of our thoughts. Recently, researchers have even been able to decode these signals into words. In this paper, we explore the feasibility of using this approach and report our early results and experiences in recording electromyogram (EMG) signals of software developers performing programming tasks. We believe, these techniques can shed light into the cognitive processes of developers and may even provide novel interactions in future programming environments.**

***Keywords*-electromyography; task assessment**

I. Introduction

As the world becomes increasing dependent on the billions and trillions lines of code written by software developers, little comfort can be taken in our limited insight into how developers manage to create software or how to educate and train new developers to meet new demands. Understanding how developers solve problems is not limited to theory-building —but can have real downstream effects in improving education, training, and the design and evaluation of tools and languages for programmers. If simple measures of cognitive effort and difficulty could be easily obtained and correlated with programming activity, then researchers could quickly identify and quantify which types of activities, segments of code, or kinds of problem solving are troublesome or improved with the introduction of a new tool.

In studying programmers, decades of psychological and observational experiments have relied on techniques such as comparing task performance, instrumenting work environments (e.g., logging key and mouse movements), or having programmers articulate their thoughts in talk-aloud protocols. Each method, when skillfully applied, can yield important insights and findings. But these methods are not without their problems. In human studies of programming, individual and task variance [1] in performance often mask any significant effects hoping to be found when evaluating a new tool. With instrumentation data, experimenters have recorded actions, but little context and must substitute cognitive measures such as cognitive effort or memory retention with metrics such as ratio of document navigations to edits or frequency of revisiting a method. Talk-aloud protocols, like surveys, rely on self-reporting and require considerable manual transcription and analysis that garner valuable but indefinite and inconsistent insight.

Within the past few decades, modern research disciplines, such as psychology and cognitive neuroscience, have collectively embraced methods that measure physiological correlates of cognition as a standard practice. One such method, electromyography (EMG), is a passive means of measuring electrical signals emitted from muscle nerves. Over a century of literature has established a deep connection between speech motor activations and cognition that can be reasonably detected with EMG.

Speculating on some possibilities: Suppose there was a method to measure every time a programmer was confused or uncertain when using a new API? Demonstratively state, a new fault localization tool or asynchronous web programming framework reduces cognitive effort. Determine how much of programming involves visual or verbal cognition? Have fine-grain measures of difficulty that predict what is time-consuming and error-prone in any programming task. Achievement of any one of these goals can have wide ramifications on how we evaluate and design tools and languages for programmers.

In this paper, we describe our efforts and early results in collecting and analyzing EMG signals acquired from programmings during two programming tasks. We provide the background, materials, proposed experiments, technical details, and emerging results for a novel method of evaluating cognitive effort when programming.

II. Background

A. Inner Speech

Inner speech is soundless mental speech that accompanies and carries our inner thoughts. During silent reading of text, we often perceive the sound of partial or complete words we encounter but make no perceivable movement of our lips or sound. However, silent reading is a relatively

1063-6897/11 $26.00 © 2011 IEEE
DOI 10.1109/ICPC.2011.49

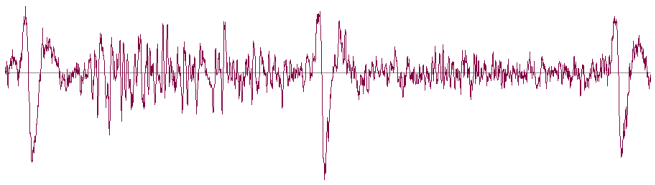

Figure 1. Raw emg signal over 2 seconds.

recent human invention: reading during medieval periods was primarily spoken aloud or in muffled tones [2]. One of the first accounts of silent reading occurs in 397 A.D. when Saint Augustine reports his astonishment of seeing his teacher, Ambrose, reading to himself [3]:

> Now, as he read, his eyes glanced over the pages and his heart searched out the sense, but his voice and tongue were silent.

But what has been understandably called silent is not necessarily so. Even for modern readers, ever so slight movement of the tongue or lips (imperceivable to the naked eye) still occurs when we read or perform mental calculations. Movement of vocal muscles itself is not necessary for thought, but merely the final recipient of motor and premotor commands sent by the brain. Dodge (1896) demonstrated this when he anesthetized his own lips and tongue and found no effect on his own inner speech [4].

Inner speech is not a perfect mirror of speech: when subjects read a novel aloud, the reading speed was 66% slower than silent reading [5]. Often a reader may only subvocalize the first part of the word.

Other late 19th and early 20th century researchers attempted to understand motor movements associated with cognition [6]. Some attempts used a glass balloon that encased the tongue to detect movement, whereas others used an inflated balloon to immobilize tongue movement. Ultimately, movement of the mouth or tongue was found to be unreliable as too much noise resulted from breathing. More success was found with readings from electromyographs (EMG) that recorded electricity from muscle nerves.

B. Subvocalization and EMG

Sokolov [6] (1972) embarked on an extensive research program that involved recording EMG during various tasks. Sokolov recorded EMG of students translating Russian literature to English and studied the effect of suppressing subvocalization during the task. Suppression resulted in reduction of correctly translated units; however, the recordings of EMG varied dramatically between individuals and even the same individuals rereading the same passage. Upon a second reading, a reader having high EMG before, may have a weak EMG or moments of no signal. Rote tasks such as recalling your name or identifying objects involved little subvocalization. One conclusion may be that people can vary the degree of subvocalization as a matter of attention and degree of study. Indeed, subvocalization has been strongly associated with rehearsal processes for memory retention and comprehension [7] and focusing attention on goals [8].

When people subvocalize, additional brain pathways are activated as a result. These activations encourage and boost hippocampal memory formation [9]. When seeing a word, not only does the meaning of the word needs to be retrieved, but the context and relevant actions associated with the word may need to be retrieved. Subvocalization activates motor and auditory pathways that extend the reach and strength of retrieval and comprehension [10].

Finally, several researchers have successfully been able to recognize words from EMG signals. An initial approach was able to recognize a set of 6 trained words with 92% accuracy [11]. More general approach based on matching phonemes and facial electrodes, has scaled to over 100 words with 10% error rate [12].

C. Alternative physiological measures

Electromyography is not the only physiological method that researchers use to study cognition. Here, we briefly highlight some of the advantages that EMG offers in contrast to other methods.

Pupillometry measures the task-evoked pupillary response over time in relation to attention. The pupil size dilates during moments of increased attention while users perform tasks. Pupillometry can be an effective measure of calculating fine-grained levels of cognitive effort, but lacks the specificity of verbal processing that subvocalization entails.

Electroencephalography (EEG) also records electrical activity, but instead from brain neurons. Intuitively, EEG may seem like a more direct method for measuring cognitive activity; however, EEG signals, due to cranial bone and skin, are an order of magnitude weaker than EMG signals and are spatially defuse. Because EEG involves very similar data collection and analysis methods, a potential technique can involve simultaneous recordings of both EEG and EMG.

fMRI measures changes in blood oxygenation levels associated with increased brain activity within 1-8 seconds to regions of brain with 1-3 mm^3 precision. Although immensely powerful, fMRI imaging requires that the measured tasks are in short duration ($<$30 seconds) and highly repeatable (100s of trials) in order to reliably detect brain activity. Not to mention, being inside a fMRI machine is not very conducive for programming tasks: participants must be placed in a small tube for a long duration with limited hand movement and extreme levels of noise.

In sum, EMG is relatively low cost but potentially high yield of cognitive measures created from a signal more reliably measured than some other techniques.

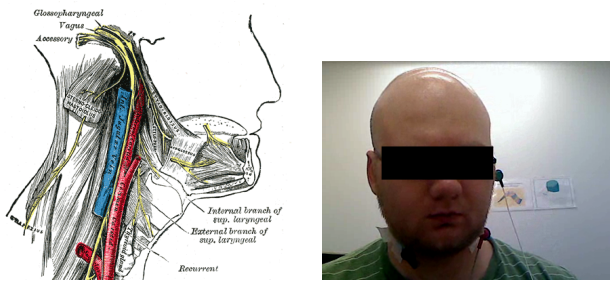

Figure 2. The vagus nerve and muscles of larynx (left) and electrodes to record larynx muscle activity (right).

III. Preliminary Experiences

A. Equipment and Target Nerves

To measure EMG, we used the Mobi EMG recording device [1], supporting up to eight channels recording at 2048 Hz. The Mobi is a light and portable device that can transmit the EMG signal in realtime via bluetooth. To synchronize the EMG signal with events from experiments or IDE instrumentation, we used a Labjack U3 [2]. Essentially, the Labjack can send digital pulses to Mobi in order mark events in the signal stream, which is necessary when correlating events such as navigation with subvocalization.

There are several muscle groups that can be targeted when measuring motor speech (see Figure 2). Facial muscles can produce strong signals when positioning the lip, and are essential for when forming certain sounds (*e.g.*, fricative phonemes in the word "fresh"). Accurately measuring tongue movements typically requires more specialized equipment. Instead, measuring the larynx, *i.e.* voice box, is more widely favored by researchers. The larynx is less likely to be activated by stray movements and does not require overt movement. However, signals from the larynx lack the richness seen in facial movements.

In our recordings, we experimented with recording 1-2 channels with dry electrodes positioned on the throat and/or near the lips (see Schultz [12] for other electrode positions).

B. Pilot Recordings

To initially test our equipment, we recorded 30 minutes of audio, video and EMG signals from silent and spoken readings of sentences from the TIMIT database. As a result, basic tool support was built for reading and displaying EMG signals and aligning with audio. One lesson learned with that it was difficult to correlate EMG with external events if there was no markers for the EMG.

To address the issue of signal segmentation, a separate channel for events was created with the help of the Labjack device. The Labjack device allowed signals to be generated and interleaved while recording EMG. Software was written to present a series of words to a subject, with events generated by Labjack prior to the word being displayed and the subject hitting a button.

After establishing the basics of recording EMG signals, focus was shifted to recording EMG simultaneously with programming tasks. A experimental workbench was developed to handle giving experimental tasks to the subject, launching an instrumented programming environment, and marking the EMG signal with the instrumented events. Two participants were given the test and data collected. After running the experiments, what was immediately apparent was that the protocol required too munch interaction with the experimenter. As a result, the protocol was modified to simplify how participants performed the tasks and interacted with the experimenter and tested on more participants.

C. Data cleaning and analysis

Any physiological recordings from humans must eliminate numerous sources of noise. Noise from heart beats and electrical devices will appear in the original raw signal (notice three heart beats in Figure 1). Such noise can typically be eliminated by using standard signal processing filter, such as a 60 Hz notch. Further, a high-pass filter can reduce noise from other sources (we used a 200 Hz,Q=0.3 high-pass filter) in our processing.

Finally, to transform the signal from continuous form into discrete events, the signal is windowed into 1/8 second chunks and passed through a low amplitude filter. Min and max value is collected and a resulting intensity is calculated.

IV. Research Questions

Research Question 1 - Can subvocalization be used to measure difficulty of programming task?

Controlled experiments often involve assigning different "difficulties" of programming tasks to software developers, but it's never clear whether they are actually more difficult for each person or why. If programmers subvocalize during programming tasks, then we may have more insight into why individuals find tasks more difficult. In future work, this question could be extended to determine which programming tools are more difficult to use or to compare difficulties in using different programming languages.

Research Question 2 - When do developers subvocalize?

Although we can record activity such as navigating or searching when a programmer codes, it is difficult to measure to what extent a programmer is actually reading and interacting with the source code text. If programmers subvocalize during some types programming activity, we may be better able to differentiate between moments of visual scanning or more involved cognitive processing. We may also be able to better differentiate between code that was relevant and irrelevant for a programming task. Future

[1] http://www.tmsi.com/?id=5
[2] http://labjack.com/u3

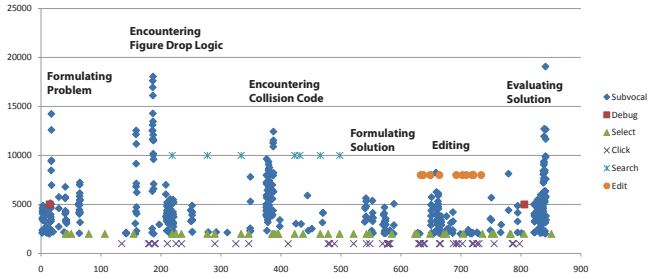

Figure 3. EMG and Programming events over 13 minutes of activity.

studies, can extend this question to probe the visual or verbal structure of cognitive processing when programming.

V. Early Results

A. Task effort

To answer research question 1, we give two programming tasks that was previously found to vary in difficulty for developers. We then measured the difference in the number and intensity of subvocalization events between the easy task and hard task. From our pilot experiments, we found a significant difference by a D-test ($p < 0.01$) in the corresponding subvocalization events and levels. These results would suggest that we can use subvocalization as an additional measure of task difficulty. However, testing with more subjects and programming tasks must be performed to confirm these results.

B. Subvocal correlates

To answer research question 2, we instrumented the activities in a programming environment and recording corresponding subvocal events during the programing task. The task was to modify Tetris to drop a falling figure completely down when pressing the space key. To gain more insight, we had the programmer describe their approach after performing the task, and then manually reviewed the recorded history to identify those activities. We then examined the correlations of the recorded history and subvocal events.

In Figure 3, we show subvocalization events (blue) and programming events collected from the task. We found subvocal activity to be strongly associated with certain activites and conditionally during others. Subvocalization was strongly associated with making edits to code. During program exploration, we found limited subvocalization. Subvocalization mainly occurred when the subject encountered important code (logic for moving Tetris block) and (testing when block stops moving). We also found subvocalization when the developer was debugging and testing the program, which may be associated with problem formulation and solution evaluation.

C. Discussion

We have observed some support for our research questions. However, more research and better standards need to be developed ensuring noise and other effects are better accounted for. Swallows, stray audio (questions) need to be automatically or manually removed to remove false events. Stray thoughts may also be an issue.

Already we have seen researchers identify subvocalized words from EMG signals. However, these are from segmented EMG signals (not continuous streams) and supervised algorithms. An alternative approach that can be used for studying programmers would be to classify characteristics of EMG signals. Can we identify introspection (when a programmer asks themselves a question), frustration (intonation), or memorization (rapid repeats that indicate attempts at rehearsal)? Future work can try unsupervised classification algorithms on labeled recordings of subvocalizations.

VI. Conclusion

The inner minds of programmers have been mostly closed to experimentation. Measuring EMG signals from subvocalization may provide a peek inside. Although, we have presented some promising results for understanding "the muscles of the mind", more research is needed. Further, each technique has different strengths and weaknesses, EMG should be considered alongside other physiological techniques and qualitative measures. Both exciting possibilities and many challenges lie ahead.

References

[1] R. E. Brooks, "Studying programmer behavior experimentally: the problems of proper methodology," *Commun. ACM*, vol. 23, pp. 207–213, April 1980.

[2] P. Saenger, *Space Between Words: The Origins of Silent Reading*. Stanford University Press, 2000.

[3] S. Augustine, *St. Augustine Confessions (Oxford World's Classics)*, H. C. (translator), Ed. Oxford Uni. Press, 1998.

[4] R. Dodge, "Die motorischen wortvorstellungen," in *Abhandlungen zur Philosophie und ihrer Geschichte*, 1896.

[5] E. Huey, *The psychology and pedagogy of reading*. Macmillian, 1908.

[6] A. N. Sokolov, *Inner speech and though*, D. B. L. (translation), Ed. Plenum Press, 1972.

[7] J. D. Smith, M. Wilson, and D. Reisberg, "The role of subvocalization in auditory imagery." *Neuropsychologia*, vol. 33, no. 11, pp. 1433–1454, November 1995.

[8] A. Miyake, "Inner speech as a retrieval aid for task goals," *Acta Psychologica*, vol. 115, no. 2-3, pp. 123–142, 2004.

[9] C. McGettigan, J. E. Warren, F. Eisner, C. R. Marshall, P. Shanmugalingam, and S. K. Scott, "Neural correlates of sublexical processing in phonological working memory." *Journal of cognitive neuroscience*, vol. 23, no. 4, pp. 961–977, April 2011.

[10] G. Hickok, K. Okada, and J. T. Serences, "Area Spt in the Human Planum Temporale Supports Sensory-Motor Integration for Speech Processing," *Journal of Neurophysiology*, vol. 101, no. 5, pp. 2725–2732, May 2009.

[11] J. C., L. D., and A. S., "Sub auditory speech recognition based on emg/epg signals," in *International Joint Conference on Neural Networks (IJCNN)*, 2003.

[12] T. Schultz and M. Wand, "Modeling coarticulation in emg-based continuous speech recognition," *Speech Commun.*, vol. 52, pp. 341–353, April 2010.