Cognitive Load Measurement from User's Linguistic Speech Features for Adaptive Interaction Design

M. Asif Khawaja[1,2], Fang Chen[1,2,3], Christine Owen[4], and Gregory Hickey[4]

[1] Nicta, ATP, Sydney, Australia
[2] School of Computer Science and Engineering, UNSW, Sydney, Australia
[3] School of Electrical Engineering and Telecommunication, UNSW, Sydney, Australia
{asif.khawaja,fang.chen}@nicta.com.au
[4] University of Tasmania, Hobart, Australia
{christine.owen,gregory.hickey}@utas.edu.au

Abstract. An adaptive interaction system, which is aware of the user's current cognitive load (CL), can change its response, presentation and flow of interaction material accordingly, to improve user's experience and performance. We present a speech content analysis approach to CL measurement, which employs users' linguistic features of speech to determine their experienced CL level. We show analyses of several linguistic features, extracted from speech of personnel working in computerized incident control rooms and involved in highly complex bushfire management tasks in Australia. We present the results of linguistic features showing significant differences between the speech from the cognitively low load and high load tasks. We also discuss how the method may be used for user interface evaluation and interaction design improvement.

Keywords: Cognitive Load, Measurement, Linguistic Features, Language usage, Word Categories, Interaction Design, Bushfire Management.

1 Introduction

Cognitive load (CL) refers to the amount of mental load imposed on a person by a particular problem-solving task. It is attributable to the limited capacity of the person's working memory and his ability to process novel information [5,6]. In complex, time-critical, and data-intense situations, users of an interaction system can experience high cognitive demands, caused either by the complexity of the task being performed or by the complex interaction designs, as in multimodal or multimedia interfaces and improper amounts of contents presented at once [7]. For example high intensity control room work-situations, such as that found in high-reliability environments e.g. air traffic control, require users or operators to manage a number of such interfaces, switching from one application interface to another, often over multiple screens and in time-critical scenarios. Operators will frequently use radios or mobile phones, make and answer calls, and speak to their co-located colleagues while completing their tasks. This can result in extremely high cognitive load and interfere with users' ability to perform at the optimum level.

Adaptive interaction systems that are aware of the users' cognitive load (CL) could in fact alleviate these problems by implementing strategies to adjust the response, presentation and flow of interaction material as per users' experienced CL to help

T. Gross et al. (Eds.): INTERACT 2009, Part I, LNCS 5726, pp. 485–489, 2009.

them complete the task effectively. However, measuring a user's CL robustly is not a trivial task. Many studies have attempted to measure CL using several methods including physiological, performance, and self-reporting subjective measures [1-3,6]. Such measures, however, can be physically or psychologically intrusive and can disrupt the normal flow of the interaction. While they may be useful approaches in research situations, they are often unsuitable for deployment in real-life scenarios.

Behavioral measures such as some speech features, e.g. pitch, prosody, pauses, and disfluencies, have also been found to be changing under high levels of CL [4,8-10]. Such measures allow non-intrusive analysis as they are based on speech data generated by users while they complete the task. Linguistic and grammatical features may also be extracted from spoken or written input for the purpose. Such features have been used before for purposes other than CL measurement [11-13].

In this paper, we present a study that analyses linguistic features of speech as potential indices of CL. We analyze several linguistic features, extracted from speech of operators working in computerized incident control rooms and involved in highly complex bushfire management tasks around Australia. We present the results of linguistic features showing significant differences between the speech from the cognitively low load and high load tasks. We also discuss how this approach may be used for user interface evaluation and interaction design improvement.

2 Study and Method

2.1 Data and Participants

Australia is one of the most bushfire-prone regions in the world. As the impact of climate change results in more extreme weather events, fire and emergency service work is becoming increasingly important and needs to be managed in order to save the communities from their effects. The speech data used in this study was collected from operators of Incident Management Teams (IMTs) [14] involved in bushfire management in Australia. Three targeted roles comprised of Incident Controller (IC), Planner, and Operations (Ops), who participated in planned bushfire management training exercises simulated to be conducted in four states of Australia (New South Wales, Victoria, Queensland, and Tasmania).

The data was collected from 9 different exercises each about 5 hours in duration on average, resulting in 27 operators' speech data being available for our study. All operators had experience in bushfire management and were competent for their assigned roles. Each exercise was monitored by a bushfire management expert training in charge. During each exercise a fire is reported that escalates and threatens local assets. The operators co-located in a control room manage the fire and communicate information and resources needed to manage it with each other as well as with the field workers and volunteers. For this purpose they use different methods e.g. phone calls, map boards, computers often with multiple screens for updated fire maps and task checklists etc. All operators' speech was recorded using lapel microphones for each exercise and was later transcribed and coded using Transana [15].

2.2 Cognitive Load Coding and Data Cleaning

All exercises were monitored by bushfire management experts who manually marked speech transcriptions for cognitive and/or task load indication based on their observations

and given subjective ratings by the operators. The transcriptions were marked for four different load levels: (1) 'low': non-bushfire activity, no time pressure; (2) 'medium': routine tasks; (3) 'high': challenging tasks, time constraints; and (4) 'very high': very challenging, lot of unexpected events and breakdowns happening. The transcribed data was later cleaned and parsed semi-automatically to bring it in a form usable by an automatic text analysis and extraction software tool. The cleaned data for each IMT role from each exercise was stored in a separate text file grouped as coded load levels resulting in 27 transcription text files.

2.3 Hypotheses

We expected several linguistic indices to be likely indicators of load including word count, negative emotions, perceptive and cognitive phrases, and inclusive words, etc. Across users, indices that we expected to increase with CL include negative emotions, number of long words, affective words (preposition and conjunction words), perceptive and cognitive phrases, and feelings and inclusive words. Indices that we expected to decrease with CL include total number of words spoken, and number of words per sentence.

3 Data Analysis and Results

During cleaning it was observed that for load levels 'low' (1) and 'very high' (4), there was insufficient data available, which could affect the results of our analysis. So to handle the problem of missing data, we combined two lower load tasks i.e. (1) and (2) into one as 'low' and two higher load tasks i.e. (3) and (4) into one as 'high' for all transcription files. These files were processed using a text analysis software called LIWC [16] that automatically extracted 85 predefined linguistic features from each transcript file for 'low' and 'high' load speech separately in that file. To take into account differences in verbosity, these were extracted as percentages of total words.

We analyzed extracted linguistic data for all three operator roles combined, as well as separately, resulting in four data sets for analysis. Tables 1 and 2 show linguistic features that showed consistent trends, i.e. either increased or decreased use of a feature between low load and high load tasks across four data sets, underscoring the importance of these features for measuring CL. The values show the usage difference in percentage for each feature, indicated by a plus sign for an increased usage and a minus sign for a decreased usage.

Table 1. Significant Linguistic Features

Features→ / ↓Data Sets	WPS	AW	NE	Per	Cog	Feel
Load-wise (All Roles)	+24%	-22%	+42%	+21%	+14%	+62%
Role-wise (IC)	+39%	-15%	+98%	+36%	+18%	+113%
Role-wise (Planning)	+22%	-48%	+80%	+20%	+4%	+20%
Role-wise (Operations)	+14%	-2%	+9%	+8%	+13%	+26%

Shaded cells = Statistically significant; $p < 0.03$

Table 2. Supporting Features

Features→ / ↓Data Sets	WC	LW	Inc
Load-wise (All Roles)	+12%	+5%	+4%
Role-wise (IC)	+4%	+13%	+5%
Role-wise (Planning)	+5%	+2%	+4%
Role-wise (Operations)	+7%	+5%	+11%

The linguistic features are listed below with few examples of each:

- WC: Total number of words spoken by the operator;
- WPS: Number of words used per sentence;
- AW: Affective words i.e. preposition and conjunction words, e.g. about, along, etc.
- NE: Words that denote negative emotions, e.g. annoy, angry, messy, afraid, etc.;
- Per: Perception words, e.g. vision, beauty, quite, rough, cold, etc.;
- Cog: Words that represent the human cognitive processes, i.e. think, consider etc.;
- Feel: Words that denote feelings, e.g. hard, difficult, heavy, loose, sharp, tight etc.
- LW: Number of long words, i.e. words with at least six letters;
- Inc: Inclusive words, e.g. and, both, each, including, plus, with etc.;

To test the significance of these features, we analyzed them for the differences between low and high load tasks for the four data sets using dependent-sample 2-tailed t-Test with 95% confidence level (alpha = 0.03 after Bonferroni adjustment). Table 1 shows test results for features with majority of them statistically significant (shaded; $p<0.03$). This implies that we can use these features to determine a user's level of CL from similar speech data robustly. Table 2 shows features that were found insignificant but may be used to support the significant features for better CL measurement due to their consistent trend across all roles.

4 Discussion

Analyses of bushfire operators speech showed consistent trends for selected linguistic features over a variety of data sets and roles, along with many significant results, and therefore, confirmed the robustness of these features for CL measurement. We remain optimistic about the lack of significant results for some roles, as this may have been due to insufficient amount of speech data. Additionally, although all operators are expected to have same language profile, some of them may not have used enough relevant words or terms for a particular linguistic feature category, due to possible difference in the nature of role.

It was interesting to find out that in contrast to our hypothesis about the WC and the WPS features, these showed increasing trends. This could be due to the complex and data-intense nature of the task. We expect same feature behavior in similar task situations but this trend may not persist with less critical task situations. Also, though the results apply across a variety of people and roles, they are specific for this combination of tasks, in a bushfire management scenario. Different linguistic features may be found to be robust for other types of application scenarios, e.g. in road or air traffic management, though it is expected to have common linguistic features across these application areas.

Adaptive interaction can be achieved with a system, which is able to determine user's experienced CL using the proposed approach. For example, in bushfire management control room scenario, the system can be able to adapt many things, from highlighting critical screen or window, to sorting and prioritizing task checklists, to showing controlled reminders, to filtering email messages, to redirecting phone calls to less cognitively loaded operators etc.

Besides the possible system adaptation, this linguistic approach to measuring CL may be used as a post-hoc analysis technique for user interface evaluation and interaction

design improvement. For example, we may evaluate two different speech-enabled interfaces to see which one is resulting in higher CL. Based on the findings we may be able to improve the interaction design for the interface causing higher CL.

5 Conclusion and Future Work

This study has provided encouraging evidence for use of linguistic features of speech as indicators of increased CL. Though these features require further cross-application validation, analysis and evaluation, they offer a promising contribution to the set of potential interactive indices that may be used by human computer interaction systems.

For future work, we intend to include in our analyses the grammatical features for CL measurement, along with validation of all the potential features and development of a common feature set for different application areas. We also intend to develop a software application to demonstrate the concept using the proposed features.

References

1. Kramer, A.F.: Physiological metrics of mental workload: a review of recent progress. In: Damos, D.L. (ed.) Multiple-task performance, pp. 279–328. Taylor and Francis, London (1991)
2. Ark, W., et al.: The Emotion Mouse. In: Bullinger, Ziegler. (eds.) HCI: Ergonomics and User Interfaces 1, vol. 1, pp. 818–823. Lawrence Erlbaum Assoc., London (1999)
3. Windell, D., Wiebe, E.N.: A Comparison of Two Mental Workload Instruments in Multimedia Instruction. In: Proc. HFES 2006. Human Factors and Ergonomics Society Press, Santa Monica (2006)
4. Yin, B., Chen, F., Ruiz, N., Ambikairajah, E.: Speech-based Cognitive Load Monitoring System. In: Proc. ICASSP 2008, pp. 2041–2044. IEEE Press, Los Alamitos (2008)
5. Chandler, P., Sweller, J.: Cognitive load theory and the format of instruction. Cognition and Instruction 8(4), 293–332 (1991)
6. Paas, F., Tuovinen, J.E., et al.: Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. Educational Psychologist 38(1), 63–71 (2003)
7. Mayer, R.E.: Multimedia learning. Cambridge University Press, Cambridge (2001)
8. Berthold, A., Jameson, A.: Interpreting Symptoms of Cognitive Load in Speech Input. In: User Modeling 1999. Springer, Wien (1999)
9. Khawaja, M.A., Ruiz, N., Chen, F.: Potential Speech Features for Cognitive Load Measurement. In: Proc. OzCHI 2007. ACM Press, New York (2007)
10. Muller, C., et al.: Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In: Proc. 8th International Conf. User Modeling, pp. 24–33 (2001)
11. Sexton, J., Helmreich, R.: Analyzing Cockpit Communication: The links between language, performance, error, and workload. University of Texas Team Research Project, Austin, USA (2000)
12. Stirman, S.W., Pennebaker, J.W.: Word Use in the Poetry of Suicidal and Nonsuicidal Poets. In: Psychosomatic Medicine, vol. 63. American Psychosomatic Society Press (2001)
13. Kramer, A., Oh, L.M., Fussell, S.R.: Using Linguistic Features to Measure Presence in Computer-Mediated Communication. In: Proc. CHI 2006. ACM Press, New York (2006)
14. Australasian Fire Authorities Council, The Australasian Inter-service Incident Management System (AIIMS), AFAC Limited, 3rd edn. (2005), http://www.afac.com.au
15. Transana, University of Wisconsin-Madison Center for Education Research, http://www.transana.org
16. Pennebaker, J.W., et al.: The Development and Psychometric Properties of LIWC 2007 (2007), http://www.liwc.net