Think before you talk: An empirical study of Relationship between Speech pauses and Cognitive load

M. Asif Khawaja
NICTA / CSE UNSW
Sydney, Australia
+61 2 9376 2157
asif.khawaja@nicta.com.au

Natalie Ruiz
NICTA / CSE UNSW
Sydney, Australia
+61 2 9376 2160
natalie.ruiz@nicta.com.au

Fang Chen
NICTA
Sydney, Australia
+61 2 9376 2101
fang.chen@nicta.com.au

ABSTRACT
Measuring a user's level of cognitive load while they are interacting with the system could offer another dimension to the development of adaptable user interfaces. High levels of cognitive load affect performance and efficiency. However, current methods of measuring cognitive load are physically intrusive and interrupt the task flow. Certain speech features have been shown to change under high levels of load and are good candidates for cognitive load indices for usability evaluation and automatic adaptation of an interface or work environment. A speech-based dual-task user study is presented in which we explore the behaviour of speech pause features in natural speech. The experiment yielded new results confirming that speech pauses are useful indicators of high load versus low load speech. We report an increase in the percentage of time spent pausing from low load to high load tasks. We interpret these results within the framework of Baddeley's modal model of working memory and detail how such a measure could be utilized in the cognitive load measurement.

Categories and Subject Descriptors
D.2.2 [Design Tools and Techniques]: User interfaces; H.5.2 [User Interfaces]: Input Devices and Strategies, Interaction Styles, Voice I/O.

General Terms
Measurement, Performance, Design, Experimentation.

Keywords
Cognitive Load, Measurement, Speech Features, Pauses.

1. BACKGROUND
This paper is an experimental extension of an earlier study by the authors [1] where, among other findings, three different speech-based feature indicators namely pause lengths, their frequencies, and latency to response were extracted from natural speech data of 24 subjects and were evaluated for any changes when users' cognitive load level was changed. Pause lengths and response latency features showed statistically significant increases between low load and high load tasks confirming the robustness of these features for cognitive load measurement. Analysis of the subjective difficulty ratings, although expected to increase, did not report any significant differences between low load and high load tasks.

In the present paper, we extend the study by increasing the total number of subjects to 48, including both male and female participants to achieve sufficient statistical power and confirm the expected significance of the subjective ratings as discussed in our previous study; normalising the speech pause differences across subjects to get more realistic pause-cognitive load relationship and/or trends; focusing on the percentages of time that users spend pausing during their natural speech, a normalised pause feature; and examining the cognitive load measurement from speech features with respect to a possible usability evaluation of speech-enabled user interfaces.

We also present the design and method of the extended user study which aims at inducing two different controlled levels of load while soliciting natural speech. We analyse the results and summarize preliminary conclusions relating to the observed behaviour.

OZCHI 2008 Proceedings ISBN: 0-9803063-4-5

2. INTRODUCTION
Cognitive load refers to the amount of mental demand imposed on a person by a particular task, and has been associated with the limited capacity of working memory and the ability to process novel information [2,3]. The cognitive load experienced by a user is derived from the semantic and representational complexity of the task being completed. High levels of cognitive load are known to decrease effectiveness and performance of users, as well as their ability to learn from their tasks [2]. Interactive user interfaces can induce high levels of cognitive load through cumbersome task representations, as in multimodal or multimedia interfaces and inappropriate amounts of content delivered at once [4].

Complex, high intensity control room work-scenarios require operators to manage a number of such interfaces, switching from one application interface to another, often over multiple screens and in time-critical situations. Operators will often use radios, make and answer phone calls, and speak to their co-located peers while completing their tasks. Research has shown that certain speech features change under high levels of load [1, 5, 6]. The vast amounts of speech data available in these environments present a rich resource for feature patterns that

betray changes in experienced cognitive load and may help in interface evaluation.

The effectiveness and usability of user interfaces has been previously assessed through physiological, performance and self-report and other subjective measures. However, such measures can be either physically or psychologically intrusive and disrupt the normal flow of the interaction, or allow only offline analyses. While they may be useful approaches in research situations, they are often unsuitable for deployment in real-life scenarios.

Particular to the assessment of cognitive load, behavioural measures, such as frequency of disfluencies or prosodic changes in speech have also been used in the art [7]. Such measures can be implicit, as they are based on the analysis of data streams employed by the users as they complete the task. These can all be standardized and allow for comparison across users [8]. Linguistic features can be extracted from spoken or written language and are highly unobtrusive. The content of the language (throughput, coherency etc) or the manner in which it is delivered (pitch, volume, articulation rate etc) can both be analysed, as well as peak intonation patterns, any of which may be indicative of high load situations.

One study has shown significant variations in the number of sentence fragments and articulation rate in users as symptoms of high cognitive load [5]. Although general trends could be observed in the relationships between the features and the cognitive load, the study did not succeed to provide a precise and accurate assessment of cognitive load. In other studies, speech variations have been shown to occur when subjects find it difficult to communicate with the system via speech, hence entering an "error-spiral" of misrecognition [9]. Such a high-load scenario causes subjects to hyper-articulate, which in turn causes changes in the speech signal [10]. At the physical end of the spectrum, the signal characteristics of speech such as energy, changes in pitch and fundamental frequency have also been explored as possible candidates for indicators of load [11]. Direct comparisons between conditions of low and high load has shown that significantly different figures for speech rate, fundamental speech frequency and speech energy could be used to distinguish between different levels of experienced load.

Traditionally in psychology, the pauses during natural speech have been associated to a person's thinking and cognitive processes, i.e. every time a person pauses during the speech, he/she processes currently known information in the limited working memory to produce the next speech response. As Schilperoord argues that if producing a response requires a particular amount of cognitive energy, then the more time it takes to produce the response, the more cognitive energy is required to do so [6]. This is another way of saying that the more you think while talking, the higher the level of cognitive load you will experience. Whether this is empirically verifiable is one of the objectives of this study.

3. USER STUDY

3.1 Dual-Task Design

The dual-task paradigm has been widely used in the field of psychology to induce high levels of cognitive load. The subject is required to perform two tasks at the same time. This becomes a much more difficult task than either one on its own. Performance is expected to degrade in each of these tasks, compared to instances where the tasks are preformed separately. This is largely due to the limited capacity of working memory as well as the load required to shift attention from one task to the other. This latter effect is known as interference [12]. Pauses are indicative of extra time-taken for problem solving and particularly in a dual-task scenario, the time it takes to manage the limited capacity of working memory as the subject works through the tasks.

In high complexity, real-time scenarios, such situations are likely to occur frequently, and users are often required to manage two or more tasks at the same time. The dual-task paradigm was chosen to help induce the high level of cognitive load in our study. The dual-task was aurally-based and consisted of playing a series of random two-digit numbers through a headset, softly in the background at random intervals, while the subject was completing a reading and comprehension task. The subjects were required to count how many numbers they heard during both reading and comprehension. The user study design choices are further discussed later in this paper.

3.2 Procedure

The subjects were asked to read aloud the selected general knowledge extracts at their own pace. They were then asked to answer a set of questions aloud and in full sentences; this speech was recorded. They did not have freedom of inspection of the extracts to answer the questions, so the contents had to be committed to memory. In the dual-task condition, the subjects were required to monitor the number count. The extracts were based on general knowledge and their difficulty level was defined using the Lexile Framework for Reading [13], which provides a standard for defining text difficulty and reading measurement by examining semantic difficulty and syntactic complexity. At the end of each reading and comprehension set, the subjects were asked to rate the difficulty of reading these extracts and again the difficulty of answering the comprehension questions on a 9-point scale, to allow us to verify whether the levels of load increased as designed.

3.3 Participants

A total of 48 subjects participated in an experiment conducted in two sessions. Session 1, with dual-task, involved 33 subjects (17 male and 16 female) while there were 15 subjects in Session 2 (7 male and 8 female). To avoid the carry-over effects, no subjects were repeated in any set. All subjects were random, remunerated and native English speakers. The differences in their reading ability would be insignificant as all subjects were adults over 18 and we assume they have relatively similar reading and comprehension abilities. The experiment was planned as a between-subjects design.

4. PRELIMINARY RESULTS

4.1 Subjective Ratings

We tested for the differences in the subjective ratings of the low and high load tasks using an independent-sample 2-tailed t-test. For the reading sub-task, average difficulty rating for the low load task was 4.2, significantly lower than that for high load task, 6.7, showing a difference of 36.4% ($t=-3.91$, $t_c=2.04$, $p<0.02$). Similarly, for the comprehension sub-task, the reported average difficulty rating for the high load task was 7.1, significantly higher than that for low load task at 4.9, with a difference of 31% ($t=-3.79$, $t_c=2.02$, $p<0.025$). This suggests

that in the dual-task situation users report an increase in their experienced level of cognitive load, as illustrated in Figure 1. This confirms the effectiveness of dual-task design.

4.2 Speech Pauses

Since the comprehension sub-task provided data most like natural speech, the analysis of pauses was carried out on this data. Analysis of pauses was conducted for two different types: silent pauses (speechless segments) and filled pauses (e.g. Ah.., hmm.., umm… etc). For each subject, we calculated percentage of total speech time the subjects spent pausing to normalize the difference across subjects. These were calculated by multiplying each subject's number of pauses by their average pause lengths and dividing the resultant by their total speech time, as presented below:

$$t = \frac{n \times l}{T}$$

where, t = percentage of time spent pausing,

n = number of pauses,

l = average pause length,

and T = Total speech time.

Percentages of time pausing (t) were calculated for silent, filled, and total pauses separately. We defined the minimum pause length of 0.3s that was taken into account. Pause lengths shorter than 0.3s were assumed to be inherent in natural speech and were not taken into consideration.

We tested for differences in percentage of time pausing in the low and high load tasks using an independent-sample 2-tailed t-test. For all kinds of pauses considered together, the percentage of time pausing for high load task (0.43) was significantly higher than that for low load task (0.30), showing a difference of 31.1% ($t=2.41$, $t_c=1.73$, $p<0.05$). This suggests a clear trend of increase in the total time pausing between the low and high load tasks as expected and is illustrated in Figure 2.

Further breaking down this data, we tested for silent pauses and filled pauses separately. For silent pauses, the percentage of time spent pausing for the low load task (0.25) was significantly lower than that for high load task (0.33), with a difference of 24.4% ($t=-2.24$, $t_c=1.71$, $p<0.05$). This confirms the increase in the silent pausing time from low to high load task which is concurring with our hypothesis and shown in Figure 2.

Unexpectedly, when testing for significant differences in the two conditions for filled pauses, none were found. This suggests that the significant difference in the percentage of total time spent pausing was actually attributed to the silent pauses and not the filled pauses.

5. DISCUSSION

Pausing in speech is a mechanism that allows the subject more time to plan, select and produce appropriate speech [6]. This time is arguably used to regulate the pace of the information flow such that the subject is able to manage their cognitive load. Analyses of the comprehension speech data were performed for selected speech pause features. Regarding the average pause lengths (l), it is worth discussing here that pauses inherently originate from breathing activity and are often very brief.

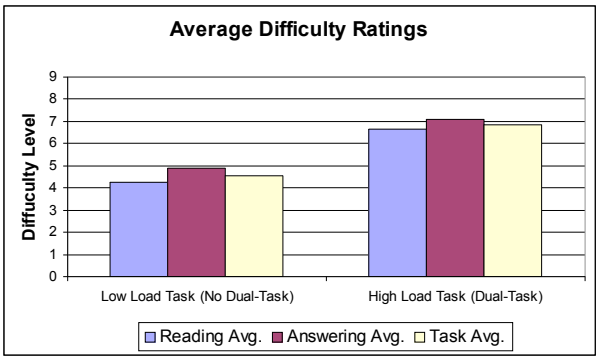

Figure 1. Subjective Ratings for Low/High Load Tasks

Therefore, to extract them from a speech database, it is important to define a cut-off value [14]. Though selected arbitrarily, it usually ranges from 0.25s to 0.3s [15]. For our study, we assumed the threshold of 0.3s.

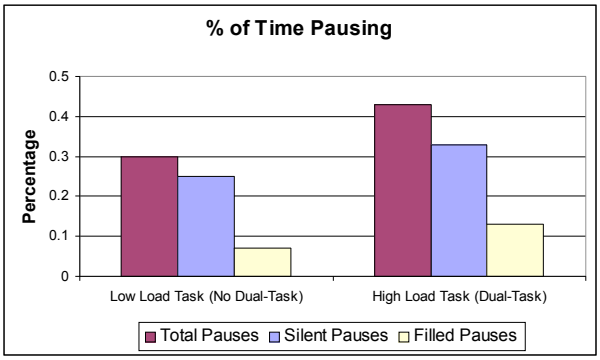

Figure 2. Percentage of Time Pausing

It has been observed that the percentage of time spent pausing increased in situations of high load. These differences are likely due to the presence of dual-task inducing higher load as in the high load task. When examining the pausing behaviour in more detail, however, we find that only silent pauses increase significantly, while the percentage of time spent on filled pauses, although showing increasing trend, does not change significantly. Also, though the results apply strongly across a wide variety of users, they are specific for this combination of tasks, in a dual-task scenario.

In this case, the primary comprehension task and the aural dual task were designed for a particular purpose. In a control room scenario as described earlier in Section 2, many tasks are completed using a visual interface – e.g. map applications, forms, and policies and contacts. However, operators are also required to complete many verbal tasks, such as handling phone calls and collecting and disseminating linguistic information from their peers. The study design was an attempt to mimic this situation artificially, in particular, the effect of this type of multi-tasking on working memory.

Using Baddeley's model of working memory [16] as a backdrop for our hypotheses, we expected that the aural dual task (linguistic comprehension and verbal production) would in fact overload the aural/verbal modal resources, namely the Phonological/Articulatory Loop, thus ensuring a high level of cognitive load in this modal area. Interpreting the results within this framework, it can be seen that when the phonological loop is overloaded, a user is less likely to emit a filled pause than a

silent pause. This may be because the execution of a filled pause itself may take up some aural/verbal resources for planning and execution in the user's working memory, and the demand cannot be accommodated due to overload, betraying fewer available resources within the user's aural/verbal working memory resources. Silent pauses, on the other hand, do not require extra processing power, and may be symptomatic of extra resources spent handling internal cognitive processes such as response planning rather than response production. This explanation may account for the way speech and in turn, speech features are affected at the physical signal or surface level, in a similar way to a real-life scenario.

The measure presented in the paper can be applied to many, data-intense, and safety-critical real-world scenarios such as air traffic control rooms, bushfire management departments, and call centres, etc. where speech is used as part of the day to day tasks, on the phone or face to face. The kind of implicit assessment available through analysis of pause durations in speech is attractive because it offers the potential of an individualised assessment of experienced load. We envisage a system that, after training, is able to detect and calculate pause durations automatically, and update these at regular intervals, such that an accurate indication of load is available at all times. If operators experiencing high load can be identified, they can be catered for with extra support if necessary, or perhaps through adaptation of the organisational or system requirements to decrease their overall experienced cognitive load to more manageable levels.

6. CONCLUSION AND FUTURE WORK

This paper presents evidence for the use of speech-based feature indicators of cognitive load, namely percentage of time spent pausing in natural speech. Though this feature requires further validation, analysis and evaluation, the results are encouraging.

Automated collection and analysis of natural speech data for relevant pauses is envisaged. This will enable interface designers and human factors to have some awareness of how the users are affected while using such interfaces, and perhaps re-designing the interface such that it reduces the overall cognitive load. We expect such individual and composite modal features to form part of a greater multimodal suite of index features acting in concert as robust indices of cognitive load.

7. ACKNOWLEDGMENTS

We thank all the subjects who participated in the experiments.

8. REFERENCES

1. Khawaja, M. A., Ruiz, N., Chen, F., "Potential Speech Features for Cognitive Load Measurement." In *Proc. OzCHI 2007*, ACM Press, 2007.
2. Chandler, P. and Sweller, J. "Cognitive load theory and the format of instruction." *Cognition and Instruction*, Vol 8, pp. 293-332.
3. Paas, F., Tuovinen, J. E., Tabbers, H., and Gerven, P. W. M. V. "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory." *Educational Psychologist*, 38, 1 (2003), 63-71.
4. Mayer, R. E., "Multimedia learning." Cambridge, UK: Cambridge University Press, 2001.
5. Berthold, A. and Jameson, A. "Interpreting Symptoms of Cognitive Load in Speech Input." In *User Modelling 99*, Springer Wien New York, 1999.
6. Schilperoord J.; "On the Cognitive Status of Pauses in Discourse Production" In T. Olive & C. M. Levy (Eds.), *Contemporary tools and techniques for studying writing*, Vol. 10, pp. 61-88. London: Kluwer Academic Publishers, 2001.
7. Muller, C., et al., "Recognizing time pressure and cognitive load on the basis of speech: An experimental study", *Proc. 8th International Conf. User Modeling,* pp.24-33, 2001.
8. Kramer, A. F., "Physiological metrics of mental workload: a review of recent progress," in *Multiple-task performance*, Damos, D. L., Ed. London: Taylor and Francis, 1991, pp. 279-328.
9. Oviatt S., "Ten myths of multimodal interaction", In *Communications of the ACM*, Vol. 42, Issue 11, pp. 74-81, 1999.
10. Oviatt, S. L., DeAngeli, A., and Kuhn, K. "Integration and synchronization of input modes during multimodal human-computer interaction." In *Human Factors in Computing Systems: CHI '97*, ACM Press, 1997.
11. Yin B., et al., "Automatic Cognitive Load Detection from Speech Features", In *Proc. OzCHI 2007,* ACM Press, 2007.
12. Kahneman, D., "Attention and effort", U.S.A: Prentice Hall, 1973.
13. Lennon C. and Burdick H.; "The Lexile Framework as an Approach for Reading Measurement and Success"; A white paper from The Lexile Framework for Reading, April 2004; http://www.Lexile.com; Last accessed: July 2008.
14. Dechert, Hans W. & Raupach, M.; "Towards a Cross-Linguistic Assessment of Speech Production." Frankfurt: Lang. 1980.
15. Schilperoord, J.; "It's about time: Temporal aspects of cognitive processes in text production."; Amsterdam / Atlanta: Rodopi, 1996.
16. Baddeley A., "Working Memory", *Science*, Vol. 255, pp. 556-559, 1992.