AUTOMATIC SPEECH FEATURE EXTRACTION FOR COGNITIVE LOAD CLASSIFICATION

Kiril Gorovoy[1], James Tung[2,3], and Pascal Poupart[2]
Department of Electrical and Computer Engineering[1], David R. Cheriton School of Computer Science[2], Department of Kinesiology[3],
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

INTRODUCTION

Performance of attention-demanding tasks is challenged if motor (e.g., walking) and cognitive (e.g., talking) tasks are carried out simultaneously. These dual-task paradigms have received increasing interest in probing the attentional influence associated with impairments to these systems [1]. For example, gait instabilities in Alzheimer's patients has been suggested to result from impaired attentional faculties impacting balance control [2]. Parkinson's patients have demonstrated inappropriate prioritization during dual-tasking, potentially leading to a higher risk of falls [3]. Furthermore, the controversy surrounding cell phone conversations on driving performance [4] further motivates the need to measure and understand the influence of attentional load.

The attentional load (or cost) placed on the central nervous system of a given cognitive task is measured by performance measures, such as reaction time or error rate. Current dual-task paradigms are limited by their capacity to measure the attentional load associated with the cognitive task. For example, a clinical test of unstable gait under dual-task conditions is failing to simultaneously perform verbal arithmetic while walking [5]. While simple to administer clinically, such tests are insensitive to mild or moderate impairments and may not account for variations in cognitive task difficulty among individuals (e.g., mathematical tasks for accountants).

More sensitive probes, such as a reaction time test, typically require 10-15 trials to provide a reliable estimate of task performance. Given endurance limitations of clinical populations and time constraints in the clinic, more efficient methods of measuring cognitive task performance are desired.

Speech Indicators of Cognitive Load

Features of speech have been identified as potential measures of performance on verbal cognitive tasks, such as verbal arithmetic. Two categories of speech features have been associated with cognitive load: i) output quality; and ii) speech output rates. Potential output quality measures include the frequency of sentence fragments, pitch, false starts, and self-repairs [6]. Output rates include measures such as articulation rate, word rate, silent pauses, and filled pauses. In comparison to output quality measures, there is stronger evidence linking rate measures to cognitive load (for review, see [6]).

Previous methods of measuring speech output rates have been limited by the need for time-intensive manual extraction [7], and often required individual baselines to account for differences in natural speaking rates [6]. Yin and Chen demonstrated the potential for automated techniques, demonstrating associations between pause and pitch peak measures to increasingly complex computer tasks [7].

The overall purpose of the current study is to develop and validate a new tool to automatically measure cognitive task performance based on speech features.

METHODS

In the current study, two specific objectives are addressed: 1) developing and validating a novel technique to automatically extract articulation rate from speech records; and 2) evaluating a multivariate approach to measure cognitive load from speech rate measures.

Speech data collection

The experiment was performed by recording speech samples of 10 undergraduate kinesiology students performing 6 tasks. Participants counted up and down by 1's, 3's, and 7's from random start numbers to simulate conditions of increasing cognitive load. It is assumed that individuals find counting by 1's easy (low cognitive load), 3's moderately difficult (intermediate difficulty), and 7's difficult (high cognitive load). Each counting exercise was performed three times, lasting approximately 20 seconds each, for a total of 18 speech samples per person. Voice data was collected using a digital voice recorder with a headset microphone.

Automating Feature Extraction

Based on reviews of the existing literature [6], pause rate, pause percentage, and articulation rate are correlated to cognitive load. While methods to extract pause rate and percentage have been previously applied [7], reliable methods to automatically estimate articulation rates are lacking.

Articulation rate, operationally defined as the number of syllables in all the uttered words divided by the total sample duration (including pauses), requires a count of the number of syllables spoken. We use a speech recognition approach to extract the number of words (and syllables) from the voice data. An open source speech to text engine (Sphinx4 [8] trained with the Wall Street Journal dictionary and a numbers (1-100) grammar) was implemented to extract the words and pauses along with their respective durations.

Syllables in a word are counted as the number of vowels, treating consecutive vowels as one, and disregarding an 'e' at the end of the word. For example the word "***o***ne" has 1 syllable and "f***i***ft***ee***n" has 2. Pauses are defined as silences lasting >100 ms, based on an optimal threshold to capture inter-sentential pauses [7]. Pause rate is calculated as the number of pauses longer than 100 ms divided by the duration of the sample. Pause percentage is the total length of all pauses greater than 100 ms divided by the total duration.

Articulation Rate Validation

One issue with speech recognition methods is a high word error rate; even words from a restricted grammar cannot be recognized with reasonable accuracy. In particular, more recognition errors are observed when the task is simple and articulation rate is high. Conversely, under greater cognitive load, verbal output tends to stretch the length of vowels or fill in pauses with filler words such as "um" or "uh". Our approach is to apply the recognizer to provide a syllable count only, not the words or content.

To evaluate the error rate in the proposed approach, we compared the approximated articulation rate with actual output rates. For each speech sample, the audio is first transcribed to determine the actual (or theoretical) articulation rate. The approximate articulation rate estimated by the speech recognizer was then compared to the actual rate using the following formula:

$$\%\ error = \frac{approximated - theoretical}{theoretical}$$

Classification procedure

To address the second objective of the study, evaluating how well the extracted features (articulation rate, pause rate, and pause percentage) relate to levels of cognitive load, we used a classification approach. Using the three speech measures for each trial as inputs, four different classification algorithms were trained and tested to distinguish the levels of cognitive load (counting by 1's, 3's, or 7's). These four algorithms (WEKA [9] implementations of the J48 decision tree, multilayer perceptron, logistic regression, and a Bayesian network), were chosen as representative of the state of the art classification methods. A cross validation method was applied where the models were trained using data from 9 participants and tested on the remaining one; alternating 10 times. Classification accuracy was measured by calculating the % of speech samples that were correctly classified as low (counting by 1's), medium (counting by 3's) and high (counting by 7's) cognitive load.

There have been previous recommendations to use a baseline articulation rate to account for individual variability in natural speaking rate [6]. To assess the potential effects of this variance, we examined the changes in classification accuracy when articulation rates were normalized to the individual mean of the 1's trials (highest rate), and 3's trials (most reflective of natural articulation rates [10]).

RESULTS

Articulation rate error

The data gathered in this study demonstrates that although speech to text recognition is poor, the number of syllables is close to the actual number. The

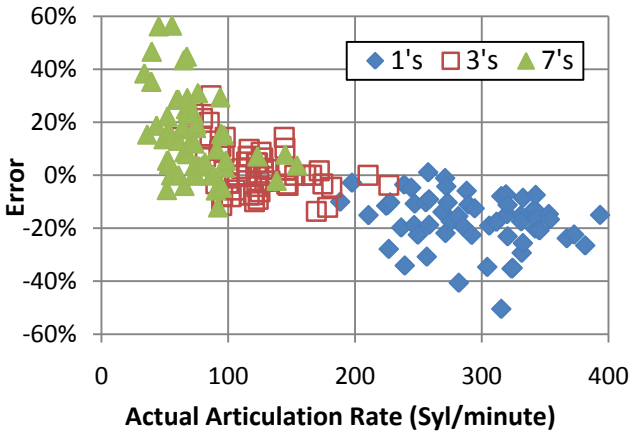
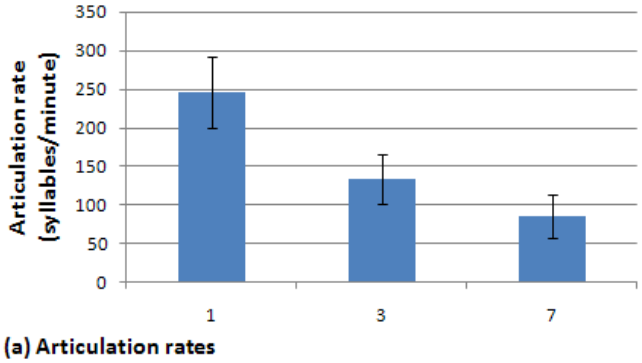
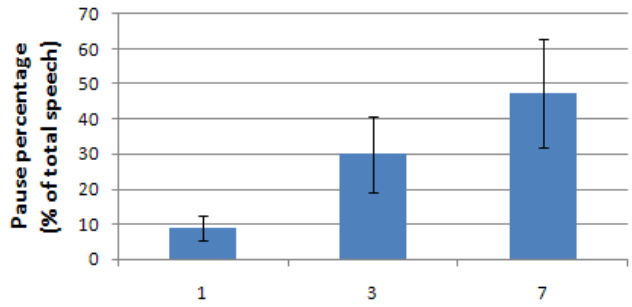


Figure 1: Percentage error vs speech rate

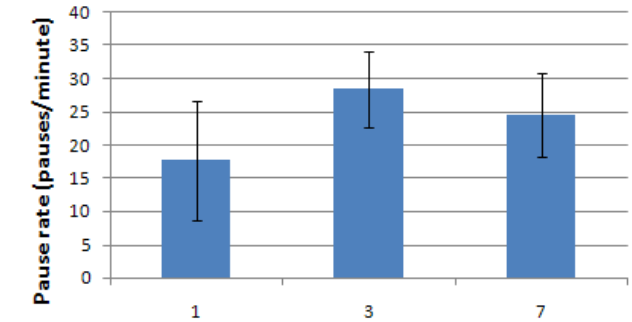
average absolute error of the approximated articulation rate is 13%. Interestingly, as shown in Figure 1, the error rate is related to the articulation rate itself, being smallest (7.1%) for the moderately difficulty task of counting by 3's, and higher (15.3% and 17.3%) for the easy and hard tasks where the participants spoke quickly or slowly, respectively. Furthermore, in the difficult task, filler words were often introduced. Despite the errors in the approximation method, the approximated articulation rate data still separates the levels of cognitive load (see Figure 2, panel a) and should not cause significant problems for classification.



(a) Articulation rates



(b) Pause percentages



(c) Pause rates

Figure 2: Speech feature trends as cognitive load increases; mean (columns) and standard deviations (error bars) are shown.

Speech feature trends

Figure 2 shows the trends of the three speech features for each counting task. Articulation rate (panel a) goes down with difficulty, while pause percentage (panel b) tends to increase as cognitive load increases. Pause rate (panel c) is low for the easy task since silences between words are often too short in duration (< 100 ms) to be considered pauses. Pause rate increases in the medium difficulty task, but decreases in the most difficult task indicating that pauses tend to be longer but fewer.

Classification Results

Table 1 shows the classification accuracy results (as % of correctly classified samples) when classifying into three levels of cognitive load using articulation rate, pause rate and pause percentage as inputs to each of the four classification algorithms. Each column compares the impact of different variants of the articulation rate. The first column shows a maximum accuracy of 85.5% using the actual (transcribed) articulation rate. The last column shows a maximum accuracy of 82.2% using the automated articulation rate technique, only a 3.3% decrease compared to using the actual articulation rate.

Table 1: Classification accuracy (% correctly classified)

Classification Algorithm	Theoretical			Approx.
	No baseline	1s as baseline	3s as baseline	No baseline
Decision Tree	83.3	76.7	81.1	77.8
Perceptron	82.2	74.4	81.1	77.8
Logistic Reg.	85.5	78.9	81.1	82.2
Bayes net	85.5	84.4	83.3	82.2

To assess the impact of variations in natural speaking rates between individuals, we examined the impact of normalizing individual speaking rates using one of the task conditions as a baseline. Previously reported articulation rates (excluding pauses) [10] include a mean articulation rate of 5.2 syl/sec for free speech, and 3.5 syl/sec in a reading task. From our results, counting by 1's (5.06 syl/sec) was the closest to natural speaking rates and was first used as a baseline to normalize individual data. As the mean articulation rate for counting by 3's (4.06 syl/sec) was closest to reported reading rates, this task was also used to normalize individual data. The impact of normalizing the observed articulation rates to these baselines is shown in the second and third columns of

Table 1. Our results indicate that neither of these baselines improved classification performance.

The accuracies reported in Table 1 are with respect to 3 levels of cognitive load corresponding to counting by 1's, 3's and 7's but not direction of counting (i.e., up or down). We also assessed the possibility of distinguishing between counting up and down (by 1's, 3's and 7's). For this finer grain 6-way classification, the best algorithm (logistic regression) had only 50.6% accuracy with the actual articulation rate and 43.9% with the approximated articulation rate.

DISCUSSION

Overall, this study validates a new technique to automatically extract the speech output performance of a verbal arithmetic task to indicate the imposed level of cognitive load. Specifically, the three speech features considered in this study, articulation rate, pause rate, and pause duration, were good indicators of cognitive demand associated with increasing levels of task difficulty. Furthermore, the examination of normalizing to individual speaking rates revealed that relative measures of (actual) articulation rates did not improve cognitive load classification accuracy.

The process of extracting the articulation rate feature can be automated despite significant speech recognition inaccuracies. Although the articulation rate error seems high at 13%, it did not have a significant impact on classification because the largest errors occurred at the extreme ranges of output rates. Our results showing the highest accuracy in the medium difficulty task where the speaker was talking at a moderate speed confirm similar observations by Siegler and Stern [11] reporting that recognition error rates were minimized at the typical utterance rates for which the recognizer was trained.

Although the individual associations between the three features and cognitive load was not examined, our results generally confirm those reported in previous work examining individual parameters [6,7], with an exception in the lack of consistent trend in pause rates. Pause rate may be useful in combination with pause percentage to indicate use of filler words as a stalling strategy in other tasks. Future work may consider extracting other features that are good indicators of cognitive load, such as mistakes and self corrections [6].

From a clinical perspective, the automated measurement of speech output rates provides a sensitive indicator of the attentional demand associated with the cognitive task component in dual-task paradigms. This approach can potentially provide a better assessment of the interaction with the motor and cognitive task demands. For example, speech output rates can be used to standardize the cognitive task difficulty in dual-task assessment of walking stability.

An eventual goal of this research is to use the proposed approach to assess the cognitive abilities of people with neuro-degenerative disease. In the current work, the automated approach was validated to cognitive load using counting tasks of varying difficulty instead of assessing participants with varying cognitive abilities. The advantage of this approach is the ability to employ within subject comparisons and minimize potential confounding factors (e.g., variable language capabilities). The next step requires evaluating the utility of our approach to accurately classify speech records of healthy participants from people with mild Alzheimer's disease in language tests such the Western Aphasia Battery [12].

ACKNOWLEDGEMENTS

This work was supported by the Alzheimer's Association and an NSERC USRA.

REFERENCES

[1] E. Fraizer and S. Mitra, "Methodological and interpretive issues in posture-cognition dual-tasking in upright stance," *Gait & Posture*, vol. 27, pp. 271–279, 2008.

[2] P.L. Sheridan, J. Solomont, N. Kowall, and J.M. Hausdorff, "Influence of executive function on locomotor function: divided attention increases gait variability in Alzheimer's disease," *Journal of the American Geriatrics Society*, vol. 51, pp. 1633–7, 2003.

[3] B.R. Bloem, Y.A.M. Grimbergen, J.G.V. Dijk, and M. Munneke, "The "posture second" strategy: a review of wrong priorities in Parkinson's disease.," *Journal of the neurological sciences*, vol. 248, pp. 196–204, 2006.

[4] I.E.H. Jr, S.M. Boss, B.M. Wise, K.E. McKenzie, and J.M. Caggiano, "Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone," *Applied Cognitive Psychology*, vol. 23, 2009.

[5] E.W. de Hoon, J.H. Allum, M.G. Carpenter, C. Salis, B.R. Bloem, M. Conzelmann, and H.A. Bischoff, "Quantitative assessment of the stops walking while talking test in the elderly," *Archives of physical medicine and rehabilitation*, vol. 84, pp. 838–842, 2003.

[6] A. Berthold and A. Jameson, "Interpreting Symptoms of Cognitive Load in Speech Input," *UM99, User modeling: Proceedings of the seventh international conference*, Springer Wien, pp. 235–244, 1999.

[7] B. Yin and F. Chen, "Towards Automatic Cognitive Load Measurement from Speech Analysis," *HCI*, pp. 1011-1020, 2007.

[8] *Sphinx-4, http://cmusphinx.sourceforge.net/sphinx4*.

[9] *Weka, http://www.cs.waikato.ac.nz/ml/weka/*.

[10] E. Jacewicz, R.A. Fox, C. O'Neill, and J. Salmons, "Articulation rate across dialect, age, and gender," *Language Variation and Change*, vol. 21, pp. 233-256, 2009.

[11] M. Siegler and R. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," *In Proceedings of ICASSP*, 1995.

[12] A. Kertesz, *The western aphasia battery*, Grune & Stratton London, 1982.