

Teaching Athletes Cognitive Skills: Detecting Cognitive Load in Speech Input

Natalie Ruiz^{^*}
^{^*}NICTA
Level 5, 13 Garden St, Eveleigh, NSW
first.last@nicta.com.au

Guang Liu^{*}

Bo Yin^{*}
[^]CSE
UNSW Kensington, NSW
first.last@cse.unsw.edu.au

Damian Farrow[~]

Fang Chen^{^*}
[~]Australian Institute of Sport
Leverrier Cres, Bruce, NSW
first.last@ausport.gov.au

As part of their preparation, athletes are often required to complete cognitive skills training using targeted sports-specific software applications. When cognitive load is very high, the quality of performance can be negatively affected and learning can be inhibited. The aim of this study is to verify whether cognitive load can be inferred directly from speech signal changes collected using one such training application. We expect that the quality of the communicative signals during interaction will change as cognitive load increases. Twelve recreational basketball players completed training requiring them to recall aloud the positions of increasing numbers of team players, and draw symbols to represent those players onto a court schematic on a digital surface. This paper focuses on the analysis of the speech data only, testing whether the speech signal changes due to high cognitive load. We describe the techniques used to build the speech load models and present the classification results. Using only automated speech signal analysis, we can identify participants experiencing low or high load with an accuracy of 92.3%. We envisage it is possible to discern broad level cognitive load ranges through speech signal changes and may provide the opportunity to tailor the training application in more appropriate ways for each learner in real time.

Cognitive load, multimodal input, speech classification.

1. INTRODUCTION

Elite athletes' performance is not only based on excellent perceptual motor abilities, but also on outstanding perceptual cognitive ability [1]. In software training applications for team sports, such as basketball for example, when athletes learn to recognise and recall strategic plays or moves from video clips or animated presentations, high levels of cognitive demand can impede learning and performance. The construct of cognitive load refers to the demands within working memory that occur during learning: too little load fails to engage learners sufficiently, whereas too much load overruns the capacity of working memory [7]. Neither condition allows effective knowledge transfer into long-term working memory and eventual automation. For this reason, being able to evaluate an athlete's cognitive demand implicitly, through observable interactive and behavioural patterns (e.g. from speech, gesture and eye-gaze) as they complete computer-based training, can provide coaches with an additional dimension for performance analysis.

2. MEASURING COGNITIVE LOAD

The first obvious indicator of high load is decreased accuracy and increased error rates. In more subtly

complex scenarios that challenge the learner, correctness-based performance is often not adversely affected. The learner can exert more effort to maintain a high level of performance should it become necessary. In such cases, response signals produced by the learner may change or become degraded in quality, due to working memory restrictions. This can occur at the content or semantic level (causing strategic changes) in responses e.g. [9], or at the signal level (such as filled pauses, false starts e.g. [2]). Monitoring the quality of the communicative signals as cognitive load increases may provide insights into the learning experience as well as the opportunity to tailor task sequences in smarter, more targeted interface adaptations. This can improve learning efficiency and maintain cognitive load within an optimal range for maximized learning outcomes.

Due to a number of factors, such as domain or interface expertise, age, or simply the time of day, learners may exhibit large individual differences when performing the same task, as result of experiencing varied cognitive load levels. Considering this, monitoring the cognitive load experienced by each learner as they learn is a key element for the development of adaptive educational user interfaces, in which the pace, content and presentation format of instruction can be dynamically adjusted to the learner's current working memory state. To date, a number of methods have been

investigated to measure cognitive load level in previous research, including: behavioural methods, e.g. text input events and mouse-click events[3], linguistic and dialogue patterns; physiological methods, such as galvanic skin response and heart rate; performance methods, such as testing and error rates; and subjective (self-report) methods of ranking experienced load level on single or multiple rating scales[1]. Among the methods, behavioural methods are probably the most suitable for practical cognitive load monitoring, which need accurate, non-intrusive, objective and online measures that can be implicitly collected as the participants interact with the software. This means that unlike performance or self-report measures, using measures derived from speech and pen interaction, would provide a much finer granularity of load indicators for subsequent content adaptation.

3. A CLASSIFICATION APPROACH

Speech features can be a particularly good choice within behavioural methods for measuring load, since speech data exists in many real domain training tasks (e.g. computer based sports training) and can be easily collected in a non-intrusive and inexpensive way. Recent research has discovered some potential features relating to cognitive load levels, such as the number of sentence fragments and articulation rate [2], and attempts to differentiate load levels from a number of high level features by using Bayesian network [3]. However, these approaches are speaker-dependent and need manually labeled data. Motivated by these limitations, we have developed the first automatic, real-time, speaker-independent cognitive load monitoring tool, details of which have been previously published [10]. The tool utilizes techniques from speech signal processing and classification research, and can be easily adapted to varied task scenarios and applications. Mel-Frequency Cepstral Coefficients (MFCC) is the de-facto standard feature in many speech recognition or classification tasks. This technique achieves highly appreciable success in comparison to other techniques specifically due to its representation of human auditory perception. Prosodic features such as pitch and intensity, on the other hand, provide extra information related to emotion or intention and have shown a potential relationship to the load levels [8]. A Gaussian Mixture Model (GMM) based classifier was implemented, with each of the load levels modelled by a GMM. The model that is the best match to a given testing sample provides the classification result during evaluation in the form of a likelihood score. Fig. 1 illustrates the process of cognitive load classification. Features extracted from raw speech have to be normalised to reduce channel and speaker variation.

Shifted Delta Coefficients are calculated to capture temporal patterns which otherwise cannot be modelled by a static model like GMM. A background model is used, represented by another GMM trained on speech data from all levels. Then, the individual load level models are adapted from it with the limited amounts of response data. Since the background model models the basic feature distribution shared by all speakers, it can be a good initial distribution for individual level models and therefore improves the precision of level models when training data is limited. To estimate the cognitive load from speech, the relevant features are extracted and enhanced, then compared to each of the level models. The model producing the highest likelihood score determines the estimated level of cognitive load.

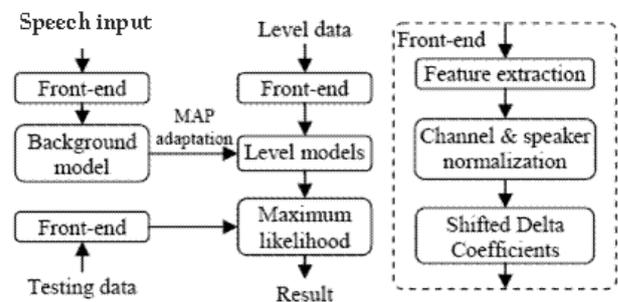


Figure 1. Process Diagram.

4. USER STUDY

The application used for this study, AISReact [5] was developed by the Australian Institute of Sport (AIS) for various types of computer-based athlete training. The application was modified to accept pen based interaction. The aim was to elicit combined speech and pen input under distinct levels of cognitive load in a real domain training task. The speech would then be used to create and test the models of classification to verify how accurately the technique can model signal changes due to high load. Likewise, the pen interaction would also be analysed for geometric and temporal changes due to cognitive load. However, in this paper, we concentrate on the analysis of speech.

4.1 Task Description

The task in the chosen training scenario consisted of a 10 second video basketball clip played on a tablet monitor. The video clip was frozen and replaced with a blank court schematic. The participants used a digital pen to mark directly on the court the locations of increasing numbers of players. There were three task playlists, for three levels of task complexity: Low, Medium and High. In the Low load task, participants needed to only remember 3 out of 10 players, in the Medium task, 6 players and finally in the High task, all 10 players. The set-up is shown in Fig.2.



Figure 2. Physical Set-Up.

4.2 Participants

Twelve male recreational basketball players, aged 19-36, each with more than 2 years experience (average of 9.41 years) volunteered to complete the study.

4.3 Materials and Procedure

Each session comprised 6 tasks (clips) for each level of complexity. The video clip footage was filmed from above, covering half the court and camera angled for a vertical orientation. All plays moved from the bottom of the screen towards the top, where the basketball hoop was located as seen in Fig. 3. The participants used a tablet monitor and digital pen to mark the appropriate number of player positions. Attacker positions were to be marked with crosses, and defender positions with circles. The ball carrier was assigned a special mark, a circle with a dot in the middle. Sample markings are shown in Fig. 4. Participants were also instructed to think-aloud through their answers, and these were captured using a close-talk microphone.



Figure 3. Basketball Video Clip End Freeze Frame.



Figure 4. Blank court image with player location markings.

5. RESULTS

5.1 Subjective Ratings and Performance

Subjective ratings were collected using a Lickert 9 point scale, where 1 was minimal effort and 9 was extreme effort. The task complexity levels induced extreme levels of load as reflected in the subjective ratings, increasing significantly as cognitive load increased, with mean averages of 3.2 (SD=1.34), 5.5 (SD=1.62) and 7.6 (SD=1.23) for the Low, Med and High load tasks respectively in Fig. 4. Due to the non-parametric dataset, this was verified using Friedman's χ^2 test ($\chi^2(12,2)=25.53, p<0.001$), where Low, Med and High were ranked 1.00, 2.04 and 2.96 respectively. As expected, the participants' performance decreased significantly from Low load to High load. Scores were given for each mark whose centroid was placed within a radius of 8% screen distance (in pixels) from the correct player position, as recommended by basketball experts at the Australian Institute of Sport, who also annotated the correct player positions on the schematic. The mean score for the Low, Med and High load tasks 83.53% (SD=11.63), 77.72% (SD=12.26) and 68.06% (SD=15.14). The decrease was verified through a repeated-measures ANOVA test ($F(2,22)=4.84, p=.018$). Subsequent planned contrasts show a significant linear ($F(1,11)=5.59, p=0.04, r = 0.46$) to the 0.05 level, with a medium effect size. This is evident in Fig. 5 also, where the performance decreases linearly first between Low and Medium load levels and then more steeply from Medium to High levels.

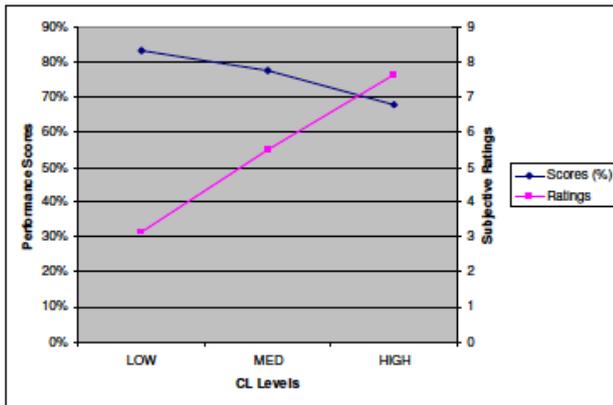


Figure 5. Performance Scores and Subjective Ratings

Overall, participants' performance decreased significantly, while their subjective ratings of load increased significantly, from Low load to High load, validating that the responses elicited by these tasks are affected by extreme levels of cognitive load.

5.2 Data Allocation and Model Training

In order to extract the effective speech from each of the samples, the non-speech parts were removed from the speech signal, leaving only the effective speech. Each treated sample was segmented into two parts of equal length. The first half was used to train cognitive load level models, while the second half was used as testing samples. Due to the nature of tasks, the amount of the participants' speech collected in the Low load tasks was significantly less than the Medium and High load tasks. In the Low load task only 3 player locations needed to be identified, with more players identified in the other levels. This results in unbalanced data allocation among classes, which introduces bias in classification. A testing sample in this case has more chance to receive a high likelihood score in a model which is over-represented. To address this, the training data for Medium and High load models were reduced to the same amount as for the Low load model. The most representative frames are retained.

	Low	Medium	High
Number of samples	63	67	65
Effective speech before (s)	276	440	590
Effective speech after (s)	276	276	276
Training speech (s)	138	138	138
Testing speech (s)	138	138	138

Table 1. Speech Samples Sizes.

5.3 Classification Results

In Table 2, the number in each cell represents the percentage of the testing samples from the class indicated by the rows, were classified to the class indicated by the columns. These numbers are the average of the two evaluation folds. It appears the

Low load pattern was captured well and achieved 100% accuracy, with all testing samples classified into the correct class. This means that in the Low level tasks all participants experienced similar low cognitive load and their speech revealed closely aligned feature patterns. A significant pattern was also captured from the speech in High load tasks across all participants, with 82% of testing samples were correctly classified into this class. Interestingly however, testing samples from the Medium load level were mostly misclassified into either the Low or High load, suggesting that no distinct pattern was captured. We suspect participants with subtly varied basketball skills and load capacity may have experienced slightly lower or higher loads in this level.

		Classified as		
		Low	Medium	High
Testing samples from	Low	100%	0%	0%
	Medium	40%	6%	54%
	High	15%	3%	82%

Table 2. The confusion matrix of three-level classification.

Observing the likelihood scores of testing samples only from the Low and High load sets, there was at least one order of magnitude raw difference between them. This gap is consistent with subject performance results showing 88.01% and 70.67% mean scores respectively. As expected, the classification with two extreme levels yielded 92.3% accuracy as shown in Table 3.

		Classified as	
		Low	High
Testing samples from	Low	100%	0%
	High	15%	85%

Table 3. The confusion matrix of two-level classification.

The results in this study provide further empirical basis to our hypothesis that the quality of the communicative signal changes when participants are confronted with high cognitive load. Our preliminary analysis of the speech data collected in this study is promising, suggesting signal-based speech features do change as cognitive load increases, irrespective of the speech contents.

6. DISCUSSION

Cognitive effort is a functional characteristic of learning [8]. The assessment of cognitive load through analysis of implicit interactive responses produced during training holds great potential for new and innovative interface adaptation strategies. While we have presented our work-in-progress results based only on speech data, yielding 92% accuracy in a two-level load classification, and we expect to find similar communicative signal

changes in pen-input. The collection and combined automated analysis of speech or pen input appears to be a logical next step, such that the modalities can triangulate to produce a more robust indication of cognitive load for training applications. The estimated cognitive load levels can also be used as an objective user feedback, to dynamically adjust the interface of training system, providing a more effective individualized learning environment. In the case of sports training, technology that can be used to diagnose underdeveloped perceptual-cognitive abilities can act as a support tool for coaches in the design of individualised training schedules, and provide athletes themselves with an additional perspective into their progress. Many computer-based training applications already exist and are routinely used in sports contexts; automated speech analysis tools can be integrated to a cognitive training tool, such as AISReact, in order to dynamically adapt the training for each athlete depending on their cognitive load level. Adaptation strategies can range from selecting task playlists based on complexity levels, or extending or collapsing the number of trials blocks to be completed, depending on a combination of performance scores and automated cognitive load analysis from speech and pen features. Finer grained changes can also be implemented in this training scenario, e.g., increasing the duration of the freeze time at the end of the clip, giving participants more time to learn recall strategies, before gradually reducing it again as cognitive load decreases and performance increases. The strategies themselves can be evaluated for each individual athlete using the same load assessment tools, adapting the interface and training schedule for maximum personalisation.

7. REFERENCES

- [1] Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., and Volpe, C. E. 1995. Defining competencies and establishing team training requirements. In R. Guzzo & E. Salas (Eds.), *Team effectiveness and decision making in organisations*. San Francisco, CA: Jossey-Bass.
- [2] Khawaja, M. A., Ruiz, N. & Chen, F. 2008 Think before you talk: An empirical study of Relationship between Speech pauses and Cognitive Load. In Proc. OzCHI'08, Cairns, Australia, pp. 335-338. Cairns: ACM Press
- [3] Lee, T.D., Swinnen, S.P. and Serrien, D.J., 1994. Cognitive effort and motor learning. *Quest* 46, pp. 328–344.
- [4] Liu, J. , Wong, C. K, and Hui, K. K. 2003. An adaptive user interface based on personalised learning. *IEEE Intelligent Systems*, 18(2):52–57.
- [5] Mackintosh, Colin. 2010. AIS React Software v.6.6, Australian Institute of Sport.
- [6] Muller, C., Grossmann-Hutter, B., Jameson, A., Rummer, R. & Wittig, F. 2001. Recognising time pressure and cognitive load on the basis of speech: An experimental study. In J. Vassileva, P. Gmytrasiewicz & M. Bauer (Eds.), *User Modeling: Proc. of the Eighth Intl. Conf.* Berlin: Springer.
- [7] Paas, F., Tuovinen, J., Tabbers, H. & Van Gerven, P. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1), 63–71.
- [8] Ruiz, N., Taib, R., Shi, Y., Choi, E., Chen, F. 2007. Using pen input features as indices of cognitive load. In Proc. 9th Int. Conf. on Multimodal Interfaces. (ICMI07), (pp 315– 318), Nagoya: ACM Press.
- [9] Ruiz, N., Taib, R. & Chen, F. Examining the redundancy of multimodal input. 2006. In Jesper Kjeldskov and Jeni Paay (Eds), Proc. Conf. Australian computer-human interaction SIG (OzCHI'06), (pp. 389-392) Sydney: ACM Press.
- [10] Yin, B., Chen, F., Ruiz, N. and Ambikairajah, E. 2008. Investigating Speech Features and Automatic Measurement of Cognitive Load. In Proc. IEEE 2008 Intl Wkshp on Multimedia Signal Processing (MMSP'08), Cairns, Australia, pp. 988-993. Cairns: ACM Press.