SPEECH-BASED COGNITIVE LOAD MONITORING SYSTEM

Bo Yin[2,1], Fang Chen[1,2,3], Natalie Ruiz[1,3], Eliathamby Ambikairajah[2,1]

[1]National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia
[2]School of Electrical Engineering and Telecommunications,
[3]School of Computer Science and Engineering,
The University of New South Wales, Sydney, NSW 2052, Australia
{bo.yin, fang.chen, natalie.ruiz}@nicta.com.au, ambi@ee.unsw.edu.au,

ABSTRACT

Monitoring cognitive load is important for the prevention of faulty errors in task-critical operations, and the development of adaptive user interfaces, to maintain productivity and efficiency in work performance. Speech, as an objective and non-intrusive measure, is a suitable method for monitoring cognitive load. Existing approaches for cognitive load monitoring are limited in speaker-dependent recognition and need manually labeled data. We propose a novel automatic, speaker-independent classification approach to monitor, in real-time, the person's cognitive load level by using speech features. In this approach, a Gaussian Mixture Model (GMM) based classifier is created with unsupervised training. Channel and speaker normalization are deployed for improving robustness. Different delta techniques are investigated for capturing temporal information. And a background model is introduced to reduce the impact of insufficient training data. The final system achieves 71.1% and 77.5% accuracy on two different tasks, each of which has three discrete cognitive load levels. This performance shows a great potential in real-world applications.

Index Terms – cognitive load, speech classification

1. INTRODUCTION

Cognitive load refers to the amount of mental demand imposed by a particular task [1], which reflects the pressure people experience in completing a task. Since cognitive load has been closely associated with the limited capacity of working memory and learning [1], it is crucial to maintain the load experienced by people within an optimal range to achieve the highest productivity. When people are overloaded, their ability of learning and performance of completing task will be negatively affected [1] resulting in faulty errors. Due to a number of factors, such as domain or interface expertise, age, mental or physical impediments, different people may be affected in different ways when performing a same task, therefore experience varied cognitive load levels. Considering this variation, monitoring the real-time cognitive load experienced by individuals is very important for developing adaptive user interfaces, in which the content and presentation can be adjusted to reduce error risk, and for critical operation environments in which an alarm can be triggered in advance.

A number of methods have been investigated to monitor (or measure) cognitive load level in previous research [1], including: behavioural methods, such as mouse speed and pressure, 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. Among the methods, behavioural methods are probably the most suitable for practical cognitive load monitoring systems which need accurate, non-intrusive, objective and online measures.

Speech features can be a particularly good choice within behavioural methods, since speech data exists in many real-life tasks (e.g. telephone conversation, voice control) and can be easily collected in a non-intrusive and inexpensive way. Recent research has discovered some potential features relating to cognitive load levels (CL levels), such as the number of sentence fragments and articulation rate [2], and tried to recognize CL levels from a number of high level features by using Bayesian network [3, 4]. However, these existing approaches are limited in speaker-dependent recognition and need manually labeled data. In this paper, an automatic, real-time, speaker-independent cognitive load monitoring system is developed utilizing techniques from speech signal processing and classification research, and can be easily adapted to varied task scenarios.

2. KEY TECHNIQUES AND SYSTEM DESIGN

To simplify the problem, we start from discrete levels instead of continuous CL level monitoring. Consequently, the monitoring problem can be seen as a classification problem, for which a number of speech classification techniques are ready for use. Considering the efficiency and the amount of data available, a GMM based classifier is proposed and the related techniques are investigated.

2.1. Speech features and temporal information

Mel-Frequency Cepstral Coefficients (MFCC) is the *de facto* standard feature in many speech recognition/classification tasks, and achieves highly appreciable success due to its better representation of human auditory response. Prosodic features such as pitch and intensity, on the other hand, give out extra information related to emotion or intension and have shown a potential relationship to the CL levels [5].

To capture the temporal information of features, three different approaches are investigated in this paper: Delta cepstrum, Acceleration (delta-delta) and Shifted Delta Cepstra. The Delta coefficient at frame ***n*** for cepstral stream $\boldsymbol{C_i}$ is calculated as:

$$\Delta C_i(n) = \sum_{k=-N}^{N} kC_i(n+k) / \sum_{k=-N}^{N} k^2 \qquad (1)$$

Acceleration is implemented by repeating Delta calculation on pre-calculated Delta coefficients. It provides the second order dynamic information of original features.

By capturing extra long-term feature patterns, Shifted Delta Coefficients (SDC) [6] has been reported to be superior than acceleration and delta in speech recognition tasks. The SDC feature vector at time ***t*** is calculated as:

$$F_{SDC}(t) = conc(c(t+iP+D) - c(t+iP-D))_{i=0\to k} \qquad (2)$$

where ***conc(.)*** means concatenating operation; D, P, k are parameters which are normally chooses to be 7, 1, 3.

The extra feature values from the above calculations are normally concatenated to the original feature vector to form a longer and enhanced feature vector containing temporal information.

2.2. Channel and speaker normalization

The consistency of speech in training and testing data is very important for statistical modeling. In a speaker-independent classification system, two major problems are speaker variation and channel mismatch. The latter is normally caused by the short-term distortions, linear channel effects and other interferences, and can be reduced by Cepstral Mean Subtraction (CMS) [7] technique which removes any fixed frequency response distortion simply by subtracting the corresponding time-averaged value over the entire speech utterance from each of the cepstral coefficients. To normalize the speaker variation, the Feature Warping [8] technique is used to map the feature distribution over an utterance to a unified distribution (Gaussian distribution in case of GMM classifier), thus reduce the variation. The warping calculation is applied on each of the feature coefficients individually, assuming different feature coefficients are independent. In this paper, the mapped value of the current feature value ***m*** is calculated over a sliding time window:

$$m = ipdf\left(\frac{N+\frac{1}{2}-R}{N}\right) \qquad (3)$$

where ***ipdf()*** is the inverse cumulative distribution function for normal distribution, ***N*** is the size of window, ***R*** is the ranking of the original value within the current window. An example of the signal distribution before and after feature warping is shown in Figure 1.

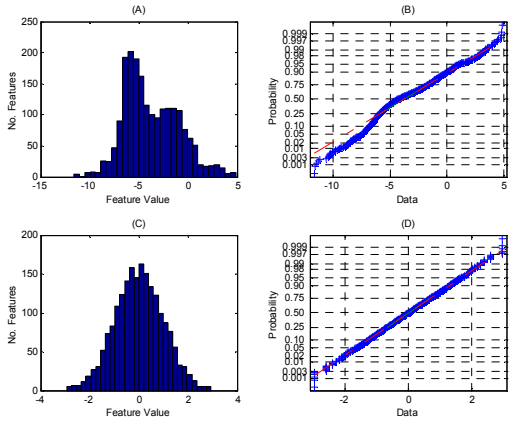

Figure 1: The distribution of a feature in a segment before (A, B) and after (C, D) warping

2.3. GMM based classifier and the background model

Considering its successful application in speech classification tasks, a GMM based classifier is proposed in this research. In this classifier, each of the CL levels was modelled by a GMM. The best matched model gives out the classification result during evaluation. According to the tasks designed for this research (described in section 3.1), the major problem to create an effective GMM classifier is the lack of training data. For example in reading task, only the answering part can be used for training level models because the models trained solely on reading data didn't show any significant difference between levels. As a solution, a background model is introduced which is another GMM trained on reading data from all levels. And then the individual CL level models are adapted from it on the limited answering data using the maximum *a posteriori* (MAP) estimation technique [9]. 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 precise of level models when training data is limited.

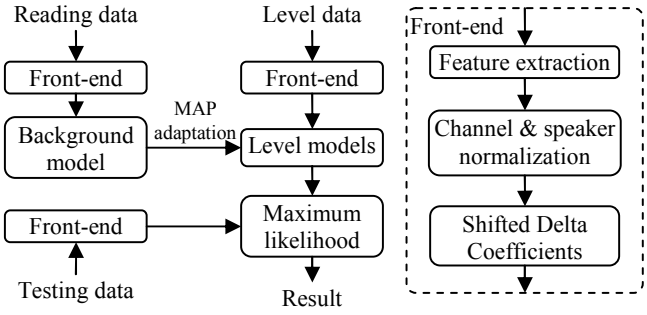

Figure 2: Diagram of the proposed monitoring system

The structure of the proposed classification system is illustrated in Figure 2. Since the evaluation process only evolves calculating and comparing likelihood scores, the classification (or monitoring) can give out result in real-time.

3. EXPERIMENTS

3.1. Task Design

To examine the hypothesis that speech features may change when a speaker experiences different cognitive loads, two different task scenarios are designed for producing the speech data by completing different tasks which induce varied loads. It is assumed the task difficulty is the major factor to influence the CL level. The proposed monitoring system is also evaluated on these two task scenarios to evaluate the performance and robustness.

3.1.1. Reading and comprehension

In this scenario, participants are required to read a short story out loudly, and then to answer three open ended questions about that story at each of the three levels. These three levels (Low/L1, Middle/L2, High/L3) contain different stories with varied difficulty level and are expected to induce different CL levels. The difficulty level of stories is measured by Lexile scale [10] – a semantic difficulty and syntactic complexity measure scale ranging from 200 to 1700 Lexiles, corresponding to the reading level expected from a first grade student to a graduate student. The stories are similar in length and contain general knowledge about weather phenomena, household appliances and the functions of the human body to avoid expertise being a factor in the results. The Lexile Ratings of the stories in L1, L2, and L3 are 925L, 1200L, and 1350L respectively.

The open ended questions are:

- Give a short summary of the story in at least five whole sentences.
- What was the most interesting point in this story?
- Describe at least two other points highlighted in this story.

Fifteen (7 male and 8 female), random, remunerated, native English speaking participants were asked to complete the reading and comprehension tasks.

3.1.2. Stroop test

The 'Stroop Test' was originally developed by John Ridley Stroop [11] for the purpose of experimental psychology research. Printed cards are prepared for the experiments with the names of colors printed with font of an incongruent color, that is, a different color than the meaning of the name. There are two types of tests: the 'Reading Color Names' (RCN), in which participants are asked to read out the words ignoring the font color; and 'Naming Colored Words' (NCW), in which the actual font color of the words has to be read out. In his research, a significant delay of task completion was noticed in NCW tests compared to RCN tests, and was explained as the automation of semantic reading interferes with the task therefore participants have to put more efforts in to override the meaning of the words to read out the actual font color. Later research conducted by Edith's group extended Stroop's tests to more situations [12], such as naming color fields, congruent color words, incongruent color words, and combined. Given the nature of these tests, they are found to be extremely useful in creating situations of different CL loads.

Six different Stroop tests are designed as three CL levels:

- Test 1 - all words are written in black.
- Test 2 - all words are written in congruent color.
- Test 3 - words are written mixed in congruent and incongruent color.
- Test 4 - words are written in incongruent color.
- Test 5 - words are written in incongruent color, appearing only one at a time.
- Test 6 - words are written in incongruent color, appearing consecutively while previous words staying on display.

RCN test 1 and 2 are used as cognitive load level 1, NCW test 3 and 4 as level 2, NCW test 5 and 6 as level 3, respectively. A separate story reading task is added for each participant to produce extra data for background model training. The set of tests were undertaken by 14, random, renumerated native English speaking participants.

3.2. Evaluation

A closed-set evaluation was conducted in case of reading and comprehension task, which means all 15 speakers appeared in evaluation data already existed in training data. For each level of the task, the reading data was collected for background training purpose, while the first two comprehension answers were used for adapting the corresponding level model and the third answer was used for evaluating the system. In average, the duration of story reading is around 90 seconds and single answer is around 30 seconds. Varied system configurations with different speech features, feature enhancement and normalization algorithms were evaluated. The number of mixtures in GMM is 256, since higher mixture numbers actually hurt due to lack of training data. The performances are shown in Table 1.

Table 1: The closed-set correction rates in various system configurations (Reading task)

System Configuration	Corr. %
MFCC	52.2%
MFCC+Prosodic (Concatenated)	59.3%
MFCC+Prosodic, Acceleration	64.4%
MFCC+Prosodic, SDC	65.7%
MFCC+Prosodic, SDC, Channel & speaker normalization	**71.1%**
MFCC+Prosodic, SDC, Channel & speaker normalization, without background model	51.1%

It is clear that prosodic features, SDC, channel & speaker normalization all significantly improve the

classification accuracy. The last configuration shows that the performance dramatically dropped if the level models are trained from corresponding reading and answering data without background model. Overall, the best performance is achieved at 71.1% in accuracy.

To investigate the system robustness and consistency on a totally different task domain, the best performed system in reading task is evaluated on the Stroop test data again.

A closed-set evaluation is conducted firstly, which was similar to the reading and comprehension task. For each one of the 14 participants, the reading data is used for training the background model while the Stroop test data is used for adapting the level models (half sets) and evaluation (the other half sets). The duration of reading is around 90 seconds and each test lasts around 30 seconds (the higher level is slightly longer than the lower one).

To investigate the performance on unknown speakers, another open-set evaluation is conducted in a 'leave-one-out' fashion. In each cycle of evaluation, a classification model is trained on the data from 13 participants and then the left one is used for evaluation. The average accuracy is given out as the open-set performance.

Table 2: The closed-set and open-set correction rates (Stroop test task)

System Configuration	Corr. %
MFCC, Prosodic, Acceleration, Channel & speaker normalization	**77.5% (closed-set)**
MFCC, Prosodic, Acceleration, Channel & speaker normalization	58.5% (open-set)

It is clear that the same classification system achieves a comparable (even higher) performance in a different task scenario. It confirms the robustness and consistency of the system. However, the performance quickly drops on unknown speakers, which means the speaker specific characteristics need to be well modeled in advance.

Table 3: Confusion matrix for closed-set stroop test

		Test results		
		L1	L2	L3
Test samples	L1	12	2	0
	L2	1	11	2
	L3	0	4	8

It can be seen from Table 3 that in all incorrectly classified instances, most utterances are misclassified into the next adjacent cognitive load level.

Compared to recent research [4] which speaker-dependently classified two cognitive load levels (high and low) in a controlled experiment with manually labeled speech features, the proposed system achieves significantly higher accuracy in much more restricted task settings without any manual interference.

4. CONCLUSION

Speech indices are some of the most promising measures for cognitive load monitoring, considering the speech features analysed are objective, can be collected in a non-intrusive way, and in many cases are already collected for communication or interaction purposes and is widely in use in many scenarios. By transforming the monitoring problem to a speech classification problem through designing tasks with discrete difficulty levels, a speech classification based cognitive load monitoring system is described in this paper. A GMM classifier based approach is proposed and several key techniques are investigated. The best-performing configuration utilized Shifted Delta Coefficients, channel and speaker normalization, and background model. Compared to existing cognitive load analysis methods, the proposed system introduces many advantages including:

- Automatic cognitive load monitoring in real-time;
- Unsupervised model training, making it easy to adapt to new scenarios without manual labelling or analysis;
- Robust for speaker-independent monitoring;

The final system achieved 71.1% and 77.5% classification accuracy in Reading and Stroop test experiments respectively, showing a great potential in real-world applications, e.g. monitoring the CL load experienced by the operators in traffic control, or by the pilots during flight simulation, in real-time.

5. REFERENCES

[1] F. Paas, J. Tuovinen, H. Tabbers, and P. V. Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational Psychologist*, vol. 38, pp. 63-71, 2003.

[2] A. Berthold and A. Jameson, "Interpreting Symptoms of Cognitive Load in Speech Input," UM99, 1999.

[3] C. M¨uller, B. Großmann-Hutter, A. Jameson, R. Rummer, and F. Wittig, "Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An Experimental Study," UM2001, 2001.

[4] A. Jameson, J. Kiefer, C. M¨uller, B. Großmann-Hutter, F. Wittig, and R. Rummer, "Assessment of a User's Time Pressure and Cognitive Load on the Basis of Features of Speech," *Journal of Computer Science and Technology*, 2006.

[5] B. Yin and F. Chen, "Towards Automatic Cognitive Load Measurement from Speech Analysis," International Conference on Human-Computer Interaction (HCII 2007), Beijing, China, 2007.

[6] B. Bielefeld, "Language identification using shifted delta cepstrum," Fourteenth Annual Speech Research Symposium, 1994.

[7] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304-1312, 1974.

[8] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," ODYSSEY-2001, 2001.

[9] S. Young, *The HTK Book*: Cambridge University Engineering Department, 2005.

[10] Metametrics, "The Lexile Framework for Reading ", 2007.

[11] J. R. Stroop, "Studies of interference in serial verbal reactions " *Journal of Experimental Psychology* 1935.

[12] D. C. Delis, J. H. Kramer, and E. Kaplan, "The Delis-Kaplan Executive Function System," *The Psychological Corporation*, 2001.