Exploring Classification Techniques in Speech based Cognitive Load Monitoring

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

The ability to monitor cognitive load level in real time is extremely useful for preventing fatal operating errors or improving the efficiency of task execution. In top of the success of our previously proposed speech based cognitive load monitoring system, we explored alternative classification techniques in this paper, including simple linear kernel Support Vector Machine (SVM), hybrid SVM-GMM which accepts the likelihood scores from GMM as inputs for SVM, and a fusion approach which integrates GMM, SVM and SVM-GMM systems together. All systems are evaluated on the data collected from two different tasks – a reading comprehension and a Stroop test based task. SVM-GMM based system achieved the highest performance on both tasks and improved the accuracy of three cognitive load levels classification from 71.1% to 75.6% and 77.5% to 82.2%, respectively.

Index Terms: cognitive load, pattern recognition, Gaussian mixture model, support vector machine, fusion

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 by us and other teams have discovered some potential features relating to cognitive load levels (CL levels), such as the number of sentence fragments and articulation rate [2], pauses and prosodic patterns [3]. Since existing approaches of recognizing high CL level such as using Bayesian network based on a number of high level features [4, 5], are limited in speaker-dependent recognition and need manually labeled data, We proposed an GMM based CL monitoring system [6], which is the first system reported to be capable to automatically classify varied CL levels from speech in real-time and speaker-independently, without any requirement of manual annotation. Later this system was improved and evaluated on two different task scenarios [7]. The results show acceptable performances on both tasks. In this paper, we investigate alternative classification techniques other than GMM, and report the performances of individual classifiers and fusion-based systems.

2. Task design

To examine the hypothesis that speech features may change when a speaker experiences different cognitive loads, controlled tasks are designed for producing speech data, under different tasks inducing varied loads. It is assumed that task difficulty is a major factor to influence CL level since it decides the intrinsic cognitive load. Two different task scenarios are designed to investigate the system robustness and consistency on totally different task domains. The actually experienced cognitive load level is validated either by performance rating or subjective rating.

2.1. 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 [8] – 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.

2.2. Stroop test

The 'Stroop Test' was originally developed by John Ridley Stroop [9] 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 [10], 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.

Three groups of Stroop tests (two tests for each) are designed with RCN tests in level low/1, NCW tests in level medium/2 and NCW plus time constraint tests in high/3. Details of these tests can be found in [7]. A reading task similar to the one in the reading and comprehension task is added before tests for collecting baseline speech data.

3. Baseline system

The baseline system is a GMM classifier based system which has achieved a success in [6, 7]. In the GMM based classifier, each of the CL levels was modelled by a GMM. The best matched model gives out the classification result during evaluation. To create an effective GMM classifier with the problem of lacking training data, a background model is introduced which is actually a GMM trained on reading data (from all levels). And then the individual CL level models are adapted from it on the corresponding answering data using the maximum *a posteriori* (MAP) estimation technique [11]. 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.

Lack of training data is not the only problem. The consistency of speech in training and testing data is also important for statistical modeling. Two major problems raised in a speaker-independent classification system 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) [12] 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 speaker variation, the Feature Warping [13] technique is used to map the feature distribution over an utterance to a unified distribution (e.g. 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.

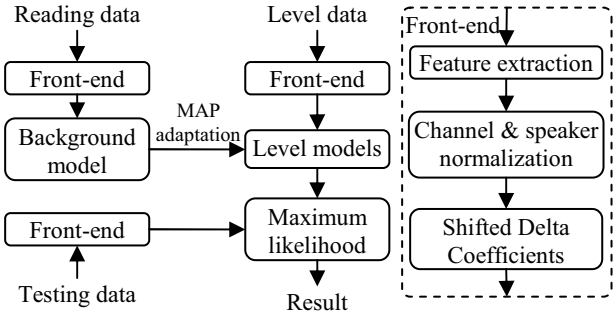

Figure 1. *Diagram of the proposed monitoring system*

The structure of the baseline system is illustrated in Figure 1. Since the evaluation process only evolves calculating and comparing likelihood scores, the classification (or monitoring) system is capable to give out result in real-time.

4. SVM classifier and fusion

To improve the performance of the existing CL monitoring system, Support Vector Machine (SVM) based classifiers are investigated in this paper. To gain the benefits from varied classifiers, fusion based systems are proposed and evaluated.

4.1. SVM and hybrid SVM-GMM classifier

SVM has received positive results in other speech recognition related areas in recent research. A typical statistical classification approach such as Gaussian Mixture Model focuses on the majority of training data and tries to model the distribution of training samples from each of the classes. Alternatively, SVM focuses on the separation edges between classes. As a hyperplane in feature space, any separation edge (separator) maximizes the distances from itself to the closest training samples from each of the classes. Therefore, these separators expose the most discriminative information. The training samples (as feature vectors) used for determining the separators are called support vectors. The feature space is separated to multiple class regions in this way. The task of training is then to find out support vectors and construct separators. Given sample $x \in X$ feature space, a two-class SVM discrimant function can be defined as [14, 15]:

$$f(x) = \sum_{k=1}^{M} \lambda_k y_k \langle x, x_k \rangle + b \qquad (1)$$

where x_k is support vector and y_k is the corresponding target class ± 1. Conventionally, only linear separators are used, which limits SVM as a linear classifier [16].

Since linear separation does not always perform well when separating samples from one class to another, a non-linear version of SVM is needed. By introducing a transforming function $\phi(.)$, Equation 1 can be written as:

$$f(x) = \sum_{k=1}^{M} \lambda_k y_k \langle \phi(x), \phi(x_k) \rangle + b \quad (2)$$

And the kernel trick can be used to avoid to evaluating $\phi(.)$. This transforming function maps the original feature space to another one, therefore effectively transforms separators to non-linear hyperplanes in the original feature space. Sometimes, this transforming process is achieved by another classifier, which is stacked onto a SVM and the output of the other classifier is used as inputs of this SVM. For example, when a GMM based classifier is stacked onto SVM, the likelihood scores produced by the GMM classifier are used as input vectors of SVM, as illustrated in Figure 2.

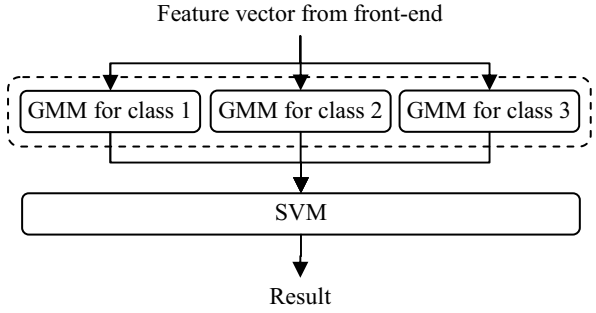

Figure 2. *A hybrid SVM-GMM classifier for 3-class classification*

As a discriminative classifier, SVM is robust to the distribution variation of the majority of training samples since the class boundaries are solely defined by those 'support vectors' on the edges.

4.2. Fusing multiple systems

There are two major different types of fusion techniques: empirical fusion such as sum-based or product-based weighting, and statistical fusion such as GMM (Gaussian Mixture Model) fusion and ANN (Artificial Neural Network) fusion.

In empirical fusion, the fused likelihood score for language hypothesis *i* is calculated either in sum or product of the scores from primary systems:

$$L_i = \sum_{j=1}^{N} w_j \cdot l_{ij} \quad (3)$$

Or

$$L_i = \prod_{j=1}^{N} l_{ij}^{w_j} \quad (4)$$

where w_j is the weighting coefficient for the likelihood scores produced by primary system j, l_{ij} is the likelihood scores produced by primary system j for language hypothesis i, and N is the total number of primary systems. In most cases an empirical process is used to optimize the weighting coefficients while occasionally performance related weightings are used, e.g. 'Matcher weighting' [17]. The final decision is based on the hypothesis with the maximum likelihood score.

If the scores are derived from probabilities, product-based weighting is usually used for combining probability and sum-based weighting is used in case of log-probabilities.

GMM based fusion is one of the popular statistical fusion techniques. In this approach, the likelihood scores produced by primary systems form a feature vector and are used to train a GMM classifier. The likelihood scores obtained from this classifier are used for final decision.

Artifical Neural Network (ANN) classifiers can also be used to fuse likelihood scores produced by primary systems [18]. With regard to the network structure, one hidden layer and one output layer has been shown to achieve reasonable performance [18]. The number of perceptrons in hidden layers and the activation functions in each layer need to be optimized on the development dataset.

As both require training from development data, GMM and ANN based fusion techniques suffer the problem that performances may deteriorate if development data is insufficient.

5. Experiments

In both task scenarios, fifteen (7 male and 8 female) native English speakers were randomly selected and remunerated as participants.

For reading and comprehension task, reading data from all levels were used for training background model, while the first two comprehension answers were used for adapting the corresponding level model from the background model and the third answer was used for evaluation. In average, the duration of effective speech for story reading is around 90 seconds, for single answer is around 30 seconds. All participants were asked to give a subjective rating regarding to each task after completion. The ratings are shown in Table 1, which scales from 1 (easiest) to 9 (hardest). A oneway ANOVA test shows the differences between levels are significant, $F(2, 87) = 62.37$, $p << .05$. Post-hoc, paired 2-tailed t-tests reveal the rating of L1 is significantly lower than L2, and L3 is significantly higher than L3.

Table 1. *Subjective rating in reading and comprehension task*

Level	Average rating (1-9)	Std. deviation
Low/1	1.57	0.60
Medium/2	3.77	3.70
High/3	6.23	3.56

For Stroop test task, similarly all reading data were used for training background model, while half of Stroop tests were used for adapting and the other half were used for evaluation. The duration of effective speech in reading is around 90 seconds and each test lasts around 30 seconds (the higher level is slightly longer than the lower one). The actually experienced cognitive load was validated by performance ratings. Task completion times in L1 is significantly less than in L2 according to a paired two-tailed t-test ($t=6.05$, $tc= 2.05$, $p<<0.05$) jumping from an average of 14.1s to 18.7s. Since time constraints were applied in L3 tests, error rates were used instead to compare the performances between L2 and L3. The average error rate jumps from 1.3% in L2 to 15.0% in L3 ($p=0.02$).

Four different systems were developed and evaluated in this paper: A GMM baseline system, a linear kernel SVM system directly accepts speech features, a hybrid SVM-GMM system, and a fusion system which combines all three previous systems. SVM implementation was based on Matlab. An empirical fusion scheme was selected instead of statistical fusion techniques because preliminary experiments suggested the performances of GMM and ANN based fusion can not match the simple empirical one, due to the amount of data available from our user studies. The same front-end as

described in [7] was used for all systems. Combined MFCC features and pitch, energy features were used and Shifted Delta Coefficients (SDC), Cepstral Mean Subtraction (CMS), Feature Warping techniques were deployed for enhancing temporal information and reducing channel/speaker variation. For all GMM based classifier, a Maximum *a-posterior* (MAP) adaptation technique is used and the number of mixtures is 256. The choices of the above configurations were based on the optimal result in previous research [7].

Table 2. *Correct classification rate of different systems in both tasks*

System	Corr. rate % in reading task	Corr. rate % in Stroop test task
GMM	71.1	77.5
SVM	66.7	77.5
SVM-GMM	**75.6**	**82.2**
FUSION	75.6	80.0

As shown in Table 2, SVM alone did not achieve a better performance than GMM based classifier. However, when stacked with GMM, the hybrid SVM-GMM system outperformed any one of the individual systems. The accuracy of three-class classification on reading task improved from 71.1% to 75.6% and from 77.5% to 82.2% on Stroop test task. An interesting observation is that fusion system only achieved a similar but not higher performance than SVM-GMM system. It may be explained as that the SVM-GMM hybrid structure already took the benefits from both SVM and GMM classifiers, while the fusion system did not provide extra discriminative information but only added in more noises.

Table 3. *Confusion matrix in Stroop test task*

		Test results		
		L1	L2	L3
Test samples	L1	13	2	0
	L2	1	13	1
	L3	0	4	11

The confusion matrix, as shown in Table 3, reveals that classification errors are mostly close errors, i.e. the segment is misclassified to the next immediate level.

6. Discussion

In this continued work, we explored varied classification techniques including linear kernel SVM, hybrid SVM-GMM, and a fusion based system. Inheriting the success of GMM based automatic cognitive load monitoring from our previous research, the proposed hybrid SVM-GMM approach further improves the performance on the same tasks. It is observed that while a classification technique targeting different discriminative region does not show a superior performance by itself, it may possibly benefit if properly integrated to existing classifiers. It is also noticed that combining multiple systems does not always contribute because the negative effects of noises may reduce the positive contribution of extra useful information. Apart from classification techniques, performance variation between tasks is interesting and needs further investigation. It may be related to the vocabulary variation between tasks. Additionally, since the current modelling technique only captures the variation pattern within the delta window (140ms in length), the information from slower variations are actually not captured. To utilize other cues relating to cognitive load, more features and modelling methods will be explored in future research.

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

[4] 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.

[5] 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.

[6] B. Yin, N. Ruiz, and F. Chen, "Automatic Cognitive Load Detection from Speech Features," OzCHI 2007, Adelaide, Australia, 2007.

[7] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based Cognitive Load Monitoring System," IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Las Vegas, 2008.

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

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

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

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

[12] 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.

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

[14] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*: Cambridge, 2004.

[15] R. e. Dehak, N. Dehak, P. Kenny, and P. Dumouchel, "Linear and Non Linear Kernel GMM SuperVector Machines for Speaker Verification," InterSpeech, 2007.

[16] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," the IEEE International Conference on Acoustic, Audio and Signal Processing, 2006.

[17] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 450-455, 2005.

[18] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language Recognition Using Phone Lattices," ICSLP, Jeju island, 2004.