Predicting cognitive load levels from speech data

Jing Su[1] and Saturnino Luz[2]

[1] Centre for Applied Data Analytics Research
University College Dublin, Ireland
jing.su@ucd.ie
[2] School of Computer Science and Statistics
Trinity College Dublin, Ireland
luzs@cs.tcd.ie

Abstract. An analysis of acoustic features for a ternary cognitive load classification task and an application of a classification boosting method to the same task are presented. The analysis is based on a data set that encompasses a rich array of acoustic features as well as electroglottographic (EGG) data. Supervised and unsupervised methods for identifying constitutive features of the data set are investigated with the ultimate goal of improving prediction. Our experiments show that the different tasks used to elicit the speech for this challenge affect the acoustic features differently in terms of their predictive power and that different feature selection methods might be necessary across these sub-tasks. The sizes of the training sets are also an important factor, as evidenced by the fact that the use of boosting combined with feature selection was enough to bring the unweighted recall scores for the Stroop tasks well above a strong support vector machine baseline.

Keywords: Paralinguistic information, cognitive load modelling, feature selection, classification.

1 Introduction

Non-verbal and paralinguistic characteristics of speech have received increasing attention from researchers. It is now commonly accepted that non-verbal sounds form an important part of human communication [2], and that non-verbal features may help identify important structural aspects of speech interaction [7] in both natural and laboratory settings [8, 1, 9]. A more recent trend in the use of paralinguistic features is their analysis for predicting levels of cognitive workload. Determination of workload levels is relevant in fields such as ergonomics, where it could help improve human computer interaction [4]. While most research in this field has been based on neurophysiological measuring, which involves specialised and intrusive equipment, the use of voice features for assessment of cognitive load levels is seen as promising enough to motivate a COMputational PARalinguistic ChallengE, ComParE [10].

This paper comprises a study of supervised and unsupervised machine learning methods applied to the prediction of cognitive load levels on a dataset distributed as part of the ComParE'14 dataset. As this dataset contains a large

number of speech and electroglottographic features, we investigated unsupervised and supervised dimensionality reduction methods in order to eliminate contingent features of the data. We then trained ensambles of classifiers (using the boosting technique) in order to distinguish among the different (discretised) levels of cognitive load.

Experiments showed that the cognitive load prediction task is better handled with supervised feature selection and different classification schemes. Contrary to our expectations, principal component analysis (PCA) feature extraction proved quite ineffective. However, with supervised feature selection a boosting global model achieved unweighted average recall (UAR) scores 20.5% and 18% higher than a published baseline based on a tuned support vector machine (SVM) classifier [10], in a Stroop time pressure and dual task, respectively. Similar per-task models were not quite as successful, but still yielded an improvement of 12% in the Stroop dual task data.

2 The Dataset

The Cognitive Load with Speech and EGG (CLSE) dataset [13, 10] was designed to support the investigation of acoustic features and evaluation of algorithms for the determination of a speaker's cognitive load and working memory during speech. The CLSE database comprises recordings of 20 male and 6 female native Australian English speakers. These recordings encompass four types of experimental tasks, namely: *reading span Sentence*, *reading span Letter*, *Stroop time pressure* and *Stroop dual task*. These tasks define four partitions of the CLSE dataset. In each case, the data instances are classified objectively into three distinct cognitive load levels: low (L1), medium (L2) and high (L3) levels.

The "span" tasks are used to measure the working memory capacity of a subject [13], in which participants are required to remember concepts or objects in the presence of distractors [3, 10]. The reading span task is based on the protocol described by Unsworth et Al. [13, 12]. It required the participants to read a series of (between two to five) possibly illogical short sentences, indicate whether the sentence read was true or false, and then remember a single letter presented briefly between sentences. This setup allowed the gatherer of the dataset to label memory load levels objectively as: L1, for data from the first sentence, L2, for data from the second sentence, and L3, for data from the third, fourth, and fifth sentences (for which no further distinctions were made).

The Stroop tasks (*Stroop time pressure* and *Stroop dual task*), named after JR Stroop's seminal experiments [11], aim to induce increased cognitive load through presentation of conflicting stimuli to the participant. In this case, the stimuli are word and colour. The participant is asked to name the font colour of words corresponding to different colour names. Data instances produced in conditions where both the colour and the word that named the colour were the same were labelled as L1 (low cognitive load). Where the font colours and the colour names differed, data were labelled L2 or L3 (medium or high level of cognitive load). The high level was defined in terms of the time pressure on

the subject (i.e. the colour had to be named in a short period of time, namely .8s) or in terms of task complexity (i.e. participants were required to perform a tone-counting task in addition to naming the font colour). These distinctions characterise the Stroop time pressure and Stroop dual task subsets of the CLSE dataset. These subsets each contain three utterances for each of three cognitive load levels per speaker.

Table 1 shows the standard "splits" of the CLSE dataset. The validation and the test set contain roughly same number of instances, while the training set contains about 50% more data. Among the four types of tasks employed in data collection, the two *span* tasks occupy the majority of the dataset while the two *Stroop* tasks comprise only about 10% of each dataset. Considering that the dataset has 6,374 attributes in total, one can readily see that the *Stroop* sets are affected more severely by high dimensionality.

Table 1. Summary of instance quantities in each type of task

	Training	Validation	Test
reading span letter	815	499	576
reading span sentence	825	525	600
stroop time pressure	99	63	72
stroop dual task	99	63	72
Total	1838	1150	1320

A fair portion of features in the training set have very low variance. This includes, for instance, all quadratic regression coefficients of level 1, and a number of other prosodic features. Some low level descriptors of spectral features also suffer from this problem. The root mean square signal frame energy feature (pcm_RMSenergy_sma_lpgain) is a case in point, with mean 1.98e-05 and variance 9.55e-10 in the training set. Such features are nearly constant and bring little discriminatory power to the classification model. We therefore removed all features with standard deviation less than 0.01. In Total 252 features (3.95% of all features) were removed from the training set, as a preprocessing step for all modelling experiments in this paper.

3 Predicting Cognitive Load Labels

A training set containing 1,838 instances described by 6,374 features challenges most classifiers since the data points are sparse with respect to dimensionality. The sparsity is more severe for models trained on subsets that contain only instances of a particular task (per task models). We therefore started by assessing the potential of two dimensionality reduction methods in rendering the dataset more tractable by learning algorithms.

3.1 PCA experiments

PCA seeks to reduce dimensionality while preserving most of data variation. Applying PCA to a dataset transformed so that all features are scaled and centered, we found that the first eight principal components explain over 95% of cumulative variance. We took 20 PCs and reencoded training and validation set into this new space. The cleaned features are projected onto the 20 PCs, and used for training (the transformed training set has 1,838 instances with 20 features). When testing with the validation set, features need to be projected to the 20 PCs before the prediction step.

Here a global model is trained and used to predict on each instance in the validation set. UAR scores were collected for each task. Contrary to our expectation, both a naive Bayes classifier and the AdaBoost classifier failed to produce satisfactory results. We found that the UAR scores were far below baseline with the SVM global model of [10]. We speculate that the reason of this low performance on the PCA-reduced sets is the lack of an effective method for normalising the data per speaker on the training and test set. In the absence of such normalisation, PCA may be dominated by a few predominant features which can easily lead this method to overfit.

3.2 Feature Selection and Global Model

Faced with the failure of an unsupervised method of dimensionality reduction, we attempted a supervised approach. The CfsSubsetEval feature filter provided by the Weka package [5] was employed. It selects attributes by individual correlation with the class variable and inter-correlation with other attributes. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred [6]. We compare global model prediction UAR scores with and without CfsSubsetEval pre-filtering in Table 2.

About classifier, we prefer Boosting with decision tree base learner instead of decision stump. The latter is a single node tree and classifies an instance by one feature. Although the feature is chosen by entropy, decision stump is too simple as a base learner in load level corpus. On the other hand, a decision tree with branching factor M=2 (minimum number of instances per leaf) by default naturally incorporates more attributes in base learner and helps the ensemble classifier.

Table 2. The effect of feature selection with AdaBoost classifier on validation set. UAR scores are from the global model, and AdaBoost is trained 30 iterations with decision tree base classifier. FS indicates feature selection with the CfsSubsetEval filter

	FS = No	FS = Yes	baseline
reading span sentence	48.50%	55.39%	61.3%
stroop time pressure	57.14%	65.08%	54.0%
stroop dual task	49.21%	52.38%	44.4%

Table 2 shows the efficacy of feature selection combined with an AdaBoost.M1 with Decision Tree base learner. Without feature selection, AdaBoostM1 beats the SVM baseline slightly in the Stroop tasks, but is 13% lower than baseline in the reading task. This observation shows the power of ensemble classification in this dataset when there is a proper base learner. When feature selection is in use, the global model achieves higher accuracy for each task. In Stroop time pressure task, the best UAR is 65.08%, an improvement of 11 points over the baseline. In the Stroop dual task, the best UAR is 52.38%, an 8-point improvement over the baseline. However, reading span is still 6% lower than baseline. In the next section we investigate per task models, where classifiers are trained on relatively more uniform training sets.

3.3 Per Task Model

In the above section, we predicted objective load level with a global model which trains a single model on all available instances and predicts on a validation set of each task. In this section we apply an alternative approach, training one model with data from one task and predicting on a validation set of the corresponding task. This is called a per task model [10]. A comprehensive training set contains objective load level instances from four tasks, part of which could be redundant for predicting on one task. Since the SVM baseline shows significantly better UAR scores with Stroop tasks, Per Task models are expected to outperform the global model in our experiments.

The split training sets are filtered in the same way as for the above described experiments. Features with standard deviation less than 0.01 are pre-filtered. The CfsSubsetEval filter selects 93, 74 and 51 features by sequence for each task. Then AdaBoost.M1 is employed as a classifier for the corresponding per task models. The number of training iterations is set to 20 for each base learner. Since the Decision Tree (DT) base learner works well for the Global model, it is used again. Moreover, we also use a Decision Stump (DS) base learner for comparison.

Table 3. The effect of feature selection with AdaBoost. UAR scores are from Per Task model, and AdaBoost is trained 20 iterations with each base learner

	Ada+DT	Ada+DS	baseline
reading span sentence	54.98%	48.86%	61.2%
stroop time pressure	68.25%	73.02%	74.6%
stroop dual task	66.67%	71.43%	63.5%

The results are shown in Table 3. Decision Stump, as the simplest tree structure, outperforms Decision Tree in AdaBoost for both Stroop tasks. This observation comes from per task model prediction on the validation set and seems quite surprising. In order to test its validity, we further analyse the Stroop Dual

Task model prediction within the training set. Figure 1a shows the performance of both DS and DT base learners under different numbers of AdaBoost iterations. It is clear that AdaBoost with the DT base learner reaches 100% UAR in the training set regardless of the number of training steps (10 to 100 iterations). At the same time, its prediction accuracy on the validation set oscillates between 61.90% and 68.25%. When we run more iterations for DT, there is no clear trend of increase or decrease in UAR on the validation set. This suggests over-fitting. In this situation, accuracy on the validation set depends on randomness of the decision boundary in the hypothesis space, and the boundary margin is already too narrow.

On the other hand, the simpler DS model improves with more training steps. Its UAR score improves in both training set and validation set when iteration increases from 10 to 20. The accuracy on the training set is far below 100%, but cannot be improved when iteration is over 20. DS reaches its upper bound of prediction power. We have seen that DS and DT both exhibit their best results on the Stroop Dual Task model, and there is no need to explore a more complex model structure. The fact that DS outperforms DT as an AdaBoost base learner is therefore to be expected. The sub-tasks with the smallest numbers of instances (Stroop dual, and Stroop time pressure) tend to favour simpler models that are less prone to overfitting.

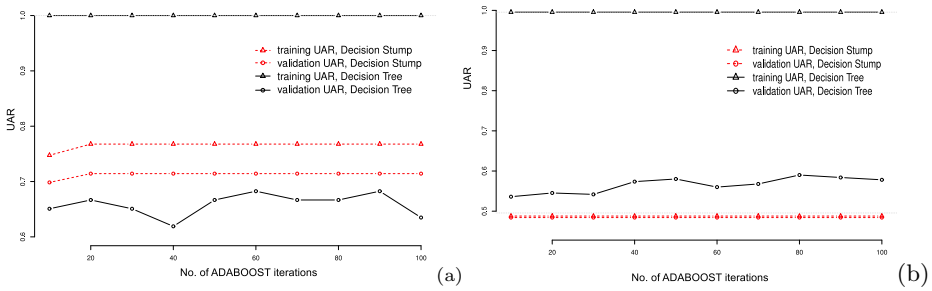

Fig. 1. Per task models of Stroop Dual Task (a) and Reading Span task (b); AdaBoost.M1 with Decision Stump and Decision Tree base learners

However, DT outperforms DS as a base learner for AdaBoost.M1 in the Reading Span Sentence task (Table 3). DS training UAR remains below 50% when training iterations increases from 10 to 100 (Figure 1b). This is a sign of under-fitting, suggesting that DS cannot represent the variances in a Reading task with 825 instances (Table 1). As in the previous per task Stroop models, the DT based classifier's training UAR is 100% when iteration equals 10, indicating that it does not suffer from the same problem. Unlike the previous case, however, in the reading task model, the UAR of DT on the validation set has a roughly increasing trend with more iterations. Prediction power is increasing with a more complex model, so here there is no indication of over-fitting. More iterations or more complex DT base learners could induce better UAR on the validation set.

4 Discussion

In this paper we proposed solutions for classifying three levels of objective load, with evidence of 6,374 speech features. In contrast to the rich feature set, there are only 1,838 instances spanning four different tasks. Since a moderately tuned SVM classifier only achieves a 44.4% baseline on a Stroop task, our results serve to emphasise the importance of data cleansing and dimensionality reduction in this modelling challenge.

In data cleansing, we dropped 252 features with standard deviation less than 0.01. These features are nearly constant, offering little value for discriminating among the three class levels while adding to the computational load. Experiments show that boosting models work well without these features, and the training time is reduced significantly. However, the number of features remaining after this pre-processing step is still very large, and dimension reduction is needed.

We found that dimensionality reduction by feature extraction through PCA harms performance in boosting as well as other models. This may be due to the differences of mean values among the features and the lack of an effective unsupervised way of normalising these values on a per speaker basis. On the other hand, the supervised CfsSubsetEval filter proved to be an effective feature selection method. The features with high correlation with class variable and low inter-correlation with other features were favoured. Multicollinearity is thus alleviated in this large feature set. The reduced feature set mainly contains frequency signals (MFCC and F0) and sound quality measures (log HNR), instead of energy related features (RMS). The reduced feature set does improve accuracy and improves on the SVM baseline for the Stroop data (Table 2).

The outcome of feature selection is encouraging, but we still need to improve model accuracy by controlling the complexity of a supervised learning model. The boosting model combines the predictions from multiple classifiers and is generally more accurate than a single classifier. The training iterations act as a controller of model complexity. In the first round, a base classifier is built. In the next round, the weight of the $n+1$ base learner is D_{n+1}, which is higher on instances that learner n has error on. The final decision is a collective vote by weighted N base learners. When boosting has no error on the training set, the generalisation power of base learner is enough for the current input. When validation accuracy keeps increasing with training accuracy stable at 100%, it is necessary to try to model with more iterations, thereby increasing the risk of over-fitting. However, when training accuracy remains stable at low values as the number of iterations increases, there is little point in proceeding. Such base learner is not complex enough to represent feature variances adequately.

5 Conclusion

We presented an exploration of feature selection and modelling trade-offs to be taken into account when approaching the challenge of categorising a speaker's cognitive load state based on acoustic features. We found that while Frequency

signals (MFCC and F0) and sound quality measures (log HNR) are critical in determining the levels of cognitive load, energy related features (RMS) seem contingent to this task.

Under appropriate settings of base learner complexity, the boosting classifier exceeds a strong SVM baseline in most Stroop tests. However, the former proved less effective in the reading span sentence tasks. This suggests that it may be necessary to study cognitive load prediction differently for each setting.

This is, however, a complex challenge and as the results reported here demonstrate, there is ample room for further exploration. In the near future we plan to investigate unsupervised ways of normalising the features per speaker as well as explore models that can take advantage of global data in per task modelling.

References

1. Bouamrane, M.M., King, D., Luz, S., Masoodian, M.: A framework for collaborative writing with recording and post-meeting retrieval capabilities. IEEE Distributed Systems Online (2004)
2. Campbell, N.: On the use of nonverbal speech sounds in human communication. In: Esposito, A., et al. (eds.) Verbal and nonverbal communication behaviours, pp. 117–128. Lecture Notes in Computer Science, Springer (2007)
3. Conway, A.R., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., Engle, R.W.: Working memory span tasks: A methodological review and users guide. Psychonomic bulletin & review 12(5), 769–786 (2005)
4. Gevins, A., Smith, M.E.: Neurophysiological measures of cognitive workload during human-computer interaction. Theoretical Issues in Ergonomics Science 4(1-2), 113–131 (Jan 2003)
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10–18 (2009)
6. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. The University of Waikato, Hamilton, New Zealand (1999)
7. Luz, S.: The non-verbal structure of patient case discussions in multidisciplinary medical team meetings. ACM Transactions on Information Systems 30(3), 17:1–17:24 (2012)
8. Luz, S., Su, J.: The relevance of timing, pauses and overlaps in dialogues: Detecting topic changes in scenario based meetings. In: Proceedings of Interspeech 2010. pp. 1369–1372. ISCA, Chiba, Japan (2010)
9. Roy, D.M., Luz, S.: Audio meeting history tool: Interactive graphical user-support for virtual audio meetings. In: Proceedings of the ESCA workshop: Accessing information in spoken audio. pp. 107–110. Cambridge University (Apr 1999)
10. Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y.: The Interspeech 2014 computational paralinguistics challenge: Cognitive & physical load. In: Proceedings of Interspeech 2014. ISCA (2014)
11. Stroop, J.R.: Studies of interference in serial verbal reactions. Journal of experimental psychology 18(6), 643–662 (1935)
12. Unsworth, N., Heitz, R.P., Schrock, J.C., Engle, R.W.: An automated version of the operation span task. Behavior research methods 37(3), 498–505 (2005)
13. Yap, T.F.: Speech production under cognitive load: Effects and classification. Ph.D. thesis, The University of New South Wales (2012)