# Prediction of Cognitive Load from Speech with the VOQAL Voice Quality Toolbox for the InterSpeech 2014 Computational Paralinguistics Challenge

*Mark Huckvale*

Speech, Hearing and Phonetic Sciences, University College London, London, U.K.

m.huckvale@ucl.ac.uk

## Abstract

This paper describes the UCL system for the cognitive load task of the Interspeech 2014 Computational Paralinguistics Challenge. The UCL system evaluates whether additional voice features computed by the VOQAL voice analysis toolbox improves performance over the baseline feature set. 144 different system configurations are evaluated on the development test set, with some systems achieving 100% classification accuracy of cognitive load in the two Stroop sub-tasks. The difficulty of the reading span sub-task is shown to be caused in part by the duration of the audio material. Performance of the best systems on the test set confirm the importance of building speaker dependent systems. While the VOQAL augmented features gave the best performance on the development test set, no benefit was found for the test set.

**Index Terms**: computational paralinguistics, cognitive load

## 1. Introduction

This paper describes the experiments performed at UCL for the Interspeech 2014 Computational Paralinguistics Challenge, specifically the cognitive-load sub-challenge. Details of the cognitive tasks, recording conditions, speech data and baseline classification performance can be found in [1]. This paper describes the motivation for our approach to the problem, the toolkits and methodology used and results on the development test set for a range of system configurations. It also provides results on the test set for the best performing systems. The paper concludes with a discussion of what has been learned about the problem of predicting speakers' cognitive load from their speech through taking part in this challenge.

## 2. Effect of Cognitive Load on Speech

Since speaking is itself a demanding cognitive task, we expect that any simultaneous task which recruits processing resources may impact the character of the planning and articulation of speech which may in turn be detectable from measurable properties of the acoustic speech signal.

Three mechanisms may be hypothesized for the effect of increased cognitive load on simultaneous speech production: (i) **Effects on planning and articulation** could arise since an increase in task complexity may make greater use of the limited resources of working memory [2] leaving less capacity for planning ahead or for constructing and executing articulator motor commands; (ii) **Effects on self-monitoring** could arise since a reduction in working memory capacity has a secondary effect of reducing perceptual attention [3], leading to "inattentional blindness", this in turn may mean that the ability of the speaker to monitor the quality of their own speech may be compromised; (iii) **Effects on articulator muscles** could occur since a reduction in working memory capacity is also known to cause an increase in activity in the autonomous nervous system [4], this in turn may lead to an increase in heart rate and muscle tone which may affect the operation of the muscles used in speaking.

These hypothetical mechanisms for how speech may be affected by a simultaneous cognitive task are likely to have different impact on the speech signal. A worsening of the ability to construct and execute motor plans might lead to changes in fluency and speaking rate. A worsening of the ability to monitor one's own speech might lead to less clear articulation. A change in the tone of muscles in the vocal tract might lead to increases in sub-glottal pressure, changes in vocal fold tension and changes in articulator position and movement. These might lead to increased vocal intensity, increased voice fundamental frequency, better glottal adduction, sharper vocal fold closures, a decrease in spectral slope and other changes in the spectral character of phonetic segments.

This analysis suggests that the effect of cognitive load on speech might operate at two levels: a high-level effect caused by changes to the planning, control and monitoring of articulator movement, and a lower-level effect caused by changes in the physiological state of the articulators. This distinction has been proposed by previous authors (e.g. [5,6]) and has led to the suggestion that a wide range of speech signal features may be required to tap into the information about cognitive load present in the speech signal. While low-level effects may be detectable by studying utterance-level distributions of voice quality, voice pitch and spectral energy, high-level effects may need features sensitive to speaking rate, dysfluency and articulatory precision.

In the Interspeech2014 Computational Paralinguistics challenge on cognitive load, speech was collected in three tasks: task 1 involved reading sentences while retaining a letter sequence in memory, while tasks 2 and 3 involved variations of the Stroop test [7] in which subjects had to read colour names from the screen or identify the font colour of colour names. We suggest that these different tasks may make different cognitive demands leading to different speech effects. The reading/memory span task made demands on memory but not on perception, while the reverse is true for the Stroop tasks. We might expect, therefore, differences in the utility of speech features for the different tasks.

The degree to which cognitive load affects speech production is also likely to be dependent on the cognitive capability of individuals. It is well know that working memory capacity varies across individuals [8], and previous studies of the effect of stressful situations on individuals show a large degree of subject variability (e.g. [9]). Thus as well as the problems of features and task described above, we should also expect considerable variability across speakers in how an increased cognitive load affects their speech.

In summary, we expect that the detection of cognitive load from speech will require a wide range of acoustic features together with a capacity to adapt to individuals.

## 3. VOQAL Toolbox

The VOQAL Toolbox is a set of speech signal analysis methods for the robust characterisation of voice quality, voice pitch and articulatory quality. The VOQAL toolbox has been developed at UCL for research into the effects of Parkinson's Disease on voice and on the effect of fatigue and cognitive load. While VOQAL shares many capabilities with the OpenSMILE toolbox [10] used for the extraction of the baseline speech features in the challenge, it has different goals. VOQAL has been designed to extract specific features related to voice quality and to do so in a robust manner. Robustness is attempted through: restricting voice quality analysis to voiced regions; voting across multiple, complementary analysis algorithms; and choosing measures which are relatively insensitive to the accurate placement of pitch marks.

VOQAL is implemented in MATLAB. Key algorithms are listed in Table 1. VOQAL builds on a number of publically available algorithms, in particular algorithms from the VOICEBOX speech analysis toolbox [11] and the GLOAT glottal analysis toolbox[12].

| VOQAL Algorithms & Features |
|---|
| **Active signal level** |
| **Syllable locations** |
| **Envelope modulation spectrogram** |
| **Voiced speech detection** |
| **Fundamental frequency (PEFAC)** |
| **Fundamental frequency (RAPT)** |
| **Fundamental frequency (SHRP)** |
| **Fundamental frequency (SWIPEP)** |
| **Glottal closure instants (SEDREAMS)** |
| **Glottal closure instants (DYPSA)** |
| **Glottal-to-Noise Excitation Ratio** |
| **Vocal Fold Excitation Ratio** |
| **Harmonic-to-Noise Ratio** |
| **Pitch Perturbation Quotient** |
| **Amplitude Perturbation Quotient** |
| **Long-Term Amplitude Spectrum** |
| **Filterbank channel energies** |
| **Filterbank channel correlations** |

Table 1. VOQAL Algorithms and Features.

For the cognitive load challenge, distributions of these parameters were collected for each utterance, and each distribution was described using the features: mean, median, standard deviation, skewness and kurtosis. Overall VOQAL provides 84 additional features per recording.

While many of these features are duplicated within the baseline OpenSMILE feature set, there is some novelty in the VOQAL features. This includes features based on the modulation spectrogram to extract changes in articulatory dynamics, filterbank channel correlations to robustly detect changes in speech effort and the vocal-fold excitation ratio measure to study the sharpness of vocal fold closures. We hope to see a gain in classification performance when the VOQAL feature set is combined with the OpenSMILE feature set for the cognitive load challenge.

## 4. Hypotheses

As well as attempting to find a good performing system for the cognitive load challenge, we also set out to test the following hypotheses that follow from the previous discussion:

1. That there will be a difference in the speech effects for the reading span task compared to the Stroop tasks. We expect to see this both in the relative difficulty of the task and in terms of the most productive features.
2. There will be strong speaker dependency effects, since the same level of cognitive load will not have the same effect on each individual and hence not the same effect on their speech.
3. The additional voice quality features provided by the VOQAL toolbox will improve performance over the OpenSMILE features alone.

## 5. Method

Systems were built for the three tasks separately: SPAN=reading/memory-span task, DUAL=Stroop dual task, TIME=Stroop time pressure task, see [1] for details.

Speaker independent and speaker dependent systems were built: ALL=speaker independent, SPKR=speaker dependent.

The baseline feature set was used, and also augmented with the VOQAL features: SMILE=baseline feature set, VOQAL=SMILE and VOQAL sets combined.

Feature normalisation was performed using a simple Z-score technique or by Gaussianisation. Normalisation was done over all speakers together or individually by speaker depending on the speaker dependency setting. Training, development and test sets were normalised separately. ZSCR=z-score method, GAUS=Guassianisation method.

Two machine learning algorithms were used: SVM = support vector machine as implemented by the R statistics [13] package `e1071` [14], CART=classification and regression tree as implemented by the R Statistics package `rpart` [15]. For the SVM based systems, a linear kernel was used and the cost parameter was set to 0.001. For the CART based systems the complexity parameter was set to 0.03.

| | | ALL | FRATIO | RELIEF |
|---|---|---|---|---|
| SMILE | SPAN | 6373 | 4210 | 3181 |
| | DUAL | 6373 | 3371 | 4153 |
| | TIME | 6373 | 4648 | 3949 |
| VOQAL | SPAN | 6457 | 4274 | 3206 |
| | DUAL | 6457 | 3410 | 4162 |
| | TIME | 6457 | 4711 | 3967 |

Table 2. Number of features selected by feature selection.

Performance using two feature selection algorithms were compared to performance using the full feature set. ALL=full feature set, FRATIO=feature selection on basis of computed F-ratios on training data, RELIEF=feature selection using the RELIEF algorithm [16]. A simple parameter threshold was

|  |  |  |  | SVM |  |  | CART |  |  |
|---|---|---|---|---|---|---|---|---|---|
| TASK | FEATURES | SPEAKERS | NORM | ALL | FRATIO | RELIEF | ALL | FRATIO | RELIEF |
| SPAN | SMILE | ALL | ZSCR | 62.9 | 61.5 | 59.6 | 47.3 | 50.4 | 47.3 |
| SPAN | SMILE | ALL | GAUS | 62.2 | **64.0** | 59.9 | 47.7 | 47.7 | 47.7 |
| SPAN | SMILE | SPKR | ZSCR | 65.8 | 66.2 | 64.7 | 51.7 | 51.7 | 48.8 |
| SPAN | SMILE | SPKR | GAUS | **66.3** | 65.7 | 63.4 | 53.6 | 53.6 | 53.6 |
| SPAN | VOQAL | ALL | ZSCR | 62.2 | 61.0 | 62.0 | 47.3 | 50.4 | 50.4 |
| SPAN | VOQAL | ALL | GAUS | 61.8 | **63.3** | 59.0 | 47.7 | 47.7 | 47.7 |
| SPAN | VOQAL | SPKR | ZSCR | 65.4 | 65.6 | 65.4 | 51.7 | 51.7 | 49.6 |
| SPAN | VOQAL | SPKR | GAUS | **67.2** | 65.4 | 65.3 | 53.6 | 53.6 | 53.6 |
| DUAL | SMILE | ALL | ZSCR | 65.1 | 66.7 | 63.5 | 61.9 | 69.8 | 69.8 |
| DUAL | SMILE | ALL | GAUS | **74.6** | 69.8 | 69.8 | 68.3 | 68.3 | 61.9 |
| DUAL | SMILE | SPKR | ZSCR | 88.9 | 90.5 | 87.3 | 71.4 | 71.4 | 71.4 |
| DUAL | SMILE | SPKR | GAUS | 90.5 | 90.5 | 93.7 | **100.0** | 100.0 | 100.0 |
| DUAL | VOQAL | ALL | ZSCR | 65.1 | 66.7 | 63.5 | 61.9 | 69.8 | 69.8 |
| DUAL | VOQAL | ALL | GAUS | **74.6** | 73.0 | 71.4 | 68.3 | 68.3 | 60.3 |
| DUAL | VOQAL | SPKR | ZSCR | 88.9 | 90.5 | 85.7 | 71.4 | 71.4 | 71.4 |
| DUAL | VOQAL | SPKR | GAUS | 88.9 | 90.5 | 92.1 | **100.0** | 100.0 | 100.0 |
| TIME | SMILE | ALL | ZSCR | 74.6 | 74.6 | 76.2 | 63.5 | 71.4 | 65.1 |
| TIME | SMILE | ALL | GAUS | **77.8** | 76.2 | 76.2 | 71.4 | 61.9 | 74.6 |
| TIME | SMILE | SPKR | ZSCR | 82.5 | 84.1 | 84.1 | 60.3 | 63.5 | 63.5 |
| TIME | SMILE | SPKR | GAUS | 84.1 | 85.7 | 87.3 | **100.0** | 100.0 | 100.0 |
| TIME | VOQAL | ALL | ZSCR | 74.6 | 74.6 | 74.6 | 63.5 | 71.4 | 65.1 |
| TIME | VOQAL | ALL | GAUS | **77.8** | 76.2 | 76.2 | 58.7 | 66.7 | 66.7 |
| TIME | VOQAL | SPKR | ZSCR | 81.0 | 84.1 | 85.7 | 60.3 | 63.5 | 63.5 |
| TIME | VOQAL | SPKR | GAUS | 84.1 | 87.3 | 85.7 | **100.0** | 100.0 | 100.0 |

Table 3. Summary of system performance. Scores are unweighted average recall percentage over three cognitive load levels. For codes see Section 5 of this paper. Systems highlighted in bold are those used to score test set.

used to select features; for the FRATIO method, features were selected if the computed F-ratio was greater than 1, for the RELIEF method, features were selected if the computed RELIEF feature weight score was greater than 0. A summary of the number of features used for the different configurations is given in Table 2.

## 6. Results

A summary of system performance on the development test set for the different configurations is shown in Table 3. Percentages are Unweighted Average Recall (UAR) based on the average percentage of each of the three cognitive load levels correctly identified by each system configuration. Notable characteristics of system performance shown in the table are the achievement of 100% identification of cognitive load levels in each Stroop task using the CART classifier.

To reveal the effect of system configuration on development set performance, a linear regression model was built between performance and the individual system configuration settings. Since it does not make sense to do this in terms of percentages, the scores are first transformed into log-odd ratios (using log-odds=log$_2$[p/1-p]). The log-odds of 100% are taken to be 6. The linear regression factors are listed in Table 4.

The regression analysis shows that the factor most affecting system performance was the choice of task. Performance was much higher on the two Stroop tasks than on the reading span task. The second most important factor is the choice of a speaker dependent system over a speaker independent system. The third most important factor was the use of Gaussianisation as a feature normalisation procedure over the use of z-scores. The choice of machine learning algorithm (SVM or CART), the choice of feature selection method (ALL or FRATIO or RELIEF), or the choice of features (SMILE or VOQAL) had no significant effect overall. However these results do not take into account interactions between factors, and the highest performing system on the SPAN task used the VOQAL features and the SVM.

The best performing systems (those shown in bold in Table 3) were then used to score the test set. Since speaker identity was not given in the test set, the test set was evaluated in two ways: firstly using the best performing speaker independent system and secondly using the best performing speaker dependent system in connection with an unsupervised speaker clustering procedure applied to the test recordings.

The speaker clustering of the test set proceeded as follows. (i) The F-ratio score for each feature in the SMILE feature set was computed with respect to speaker identity in the training set to establish the best features for speaker clustering. (ii) The best features were then selected from the test set and used to cluster the feature vectors into 8 speakers using k-means clustering. (iii) Cluster membership was then used to perform speaker-dependent normalisation of the test data.

Performance of the best-performing system configurations on the test set are shown in Table 5.

| Parameter | Factor Change | Effect size (log-odds) |
|---|---|---|
| Task | SPAN → DUAL | 1.83 |
| Task | SPAN → TIME | 1.71 |
| Speaker dependence | ALL → SPKR | 1.39 |
| Normalisation | ZSCR → GAUS | 0.94 |
| ML Algorithm | CART → SVM | 0.16 |
| Feature selection | ALL → FRATIO | 0.08 |
| Feature selection | ALL → RELIEF | 0.02 |
| Feature set | SMILE → VOQAL | -0.02 |

Table 4. Regression analysis of system configuration. The configuration change in the middle column caused the average performance change in the right hand column. Performance change expressed in log-odds.

| Features | System Type | Development | Test |
|---|---|---|---|
| SMILE | Speaker Independent | 66.5 | 63.1 |
| | Speaker Dependent | 73.2 | 57.2 |
| VOQAL | Speaker Independent | 66.1 | 62.7 |
| | Speaker Dependent | 73.9 | 55.2 |

Table 5. Performance of best-performing systems on development set and test set. Values are unweighted average recall on all sub-tasks.

## 7. Discussion

### 7.1. Task Differences

In terms of the different tasks, it is clear that the reading span task was much more difficult than either of the Stroop tasks. For comparable systems, error rates were often twice as large on the reading span task. One possible explanation for this might have been because of the difference in length of audio materials. For the reading span task each file was about 4s, while for the Stroop tasks each file was 15-20s. To investigate whether this difference affected performance, a test was made in which the training and development set audio files for the reading span task were batched into blocks of four sentences. The SMILE feature set was then recomputed for the batched audio files. Performance on the development set improved significantly. For example in the speaker independent configuration with gaussian normalisation, F-ratio features, and an SVM, UAR increased from 64.0% on the original files to 71.8% on the batched files. For the equivalent speaker-dependent system, performance increased from 65.7% to 75.4%. Since the test data could not be batched into longer chunks, it was not possible to find performance on the test set.

Another means to explore differences across tasks is to compare the most preferred features found by the feature selection process. Using the SMILE data, the 6373 features were ranked by computing F-ratios against the 3 load levels separately for each task. Of the top 1000 features selected for each task, only 30% on average were common between any pair of tasks.

### 7.2. Speaker Differences

The comparison between development set performance and test set performance shown in Table 5 highlights the importance of having speaker dependent systems. The speaker independent systems had broadly similar performance on both sets. The speaker dependent system had good performance on the development set where correct speaker identity was known, but had poor performance on the test set, where speaker identity could only be judged by unsupervised clustering.

### 7.3. Feature Set Differences

Broadly speaking the choice of feature sets and choice of feature selection process had little effect on system performance. A general observation was that performance increased as more features were included, despite the fact that the features were likely to be highly correlated. One explanation for this may be because the linear kernel used for the SVM operated better in a higher dimension space. The addition of VOQAL features to the baseline openSMILE feature set made little difference, although the single highest performing system did exploit the VOQAL features.

## 8. Conclusions

Performance on the cognitive load task could have been improved by better corpus design: we have shown how longer audio samples for the read speech task would have helped, and also that test-set performance seems to be crucially weakened by a failure to include speaker identity.

The need for a speaker dependent approach highlighted here may not relate just to individual differences in how speakers convey cognitive load in speech, but also in terms of how well individuals cope with a complex task. The load levels used in the corpus come from the complexity of the task, not from some independent measure of the cognitive loading of the subjects. Ideally, a separate physiological measure of cognitive load would be used to provide the 'gold standard' labels. One approach to this would be to make simultaneous pupillometry recordings, since changes in the autonomous nervous system as a consequence of working memory capacity limitations also suppresses parasympathetic nervous activity and causes pupil dilation [17].

## 9. Acknowledgements

Thanks to the organisers of the Interspeech 2014 Computational Paralinguistics Challenge for making this study possible. I have learned a lot through taking part.

# 10. References

[1] Schuller, B., Steidl, A., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y., "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive and Physical Load", Proc. Interspeech 2014.

[2] Barrouillet P, Bernardin S, Camos V., "Time constraints and resource sharing in adults' working memory spans". Journal of Experimental Psychology: General, 133 (2004) 83–100.

[3] Lavie, N., Hirst, A., de Fockert, J., Viding, E., "Load Theory of Selective Attention and Cognitive Control", Journal of experimental Psychology: General, 133 (2004) 339-354.

[4] Hansen, A., Johnsen, B., Thayer, J., "Vagal influence on working memory and attention", Internatioal Journal of Psychophysiology, 48 (2003) 263-274.

[5] Yap, T., "Speech production under cognitive load", PhD Dissertation, University of New South Wales, Australia, 2012. Downloaded from: http://unsworks.unsw.edu.au/fapi/datastream/unsworks:10507/SOURCE01

[6] Le, P. N., "The use of spectral information in the development of novel techniques for speech-based cognitive load classification", PhD dissertation, University of New South Wales, Australia, 2012. Downloaded from: http://unsworks.unsw.edu.au/fapi/datastream/unsworks:10313/SOURCE01

[7] Ridley, J. "Studies of interference in serial verbal reactions". Journal of Experimental Psychology 18 (1935) 643–662.

[8] Just, M., Carpenter, P., "A capacity theory of comprehension, individual differences in working memory", Psychological Review, 99 (1992) 122-149.

[9] Johannes, B., Salnitski, V.P., Gunga, H-C. and Kirsch, K., "Voice stress monitoring in space. Possibilities and limits", Aviation, Space, and Environmental Medicine 71 (2000) 58.

[10] Eyben, F., Weninger, F., Woellmer, M., Schuller, B., "openSmile: the Munich versatile and fast open-source audio feature extractor", Downloaded from http://opensmile.sourceforge.net/

[11] Brooks, M., "VOICEBOX: speech processing toolbox for MATLAB", Downloaded from: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[12] Drugman, T., "GLOAT: Glottal analysis toolbox", Downloaded from http://tcts.fpms.ac.be/~drugman/Toolbox/

[13] "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. Downloaded from http://www.R-project.org/.

[14] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, A., "e1071: Misc Functions of the Department of Statistics TU Wien", 2014. Downloaded from http://CRAN.R-project.org/package=e1071

[15] Therneau, T., Atkinson, B., Ripley, B. , "rpart: Recursive Partitioning and Regression Trees", 2104. Downloaded from http://CRAN.R-project.org/package=rpart

[16] Kira, K., Rendell, L., "The feature-selection problem: traditional methods and a new algorithm", Proc. AAAI-92, 1992, 129-134.

[17] Kahneman, D., & Beatty, J., "Pupil diameter and load on memory", Science 154 (1966) 1583–1585.