The Prediction of Fatigue Using Speech as a Biosignal

Khan Baykaner[1], Mark Huckvale[1], Iya Whiteley[2], Oleg Ryumin[3], Svetlana Andreeva[3]

[1]Speech Hearing and Phonetic Sciences, UCL, London, UK
[2]Centre for Space Medicine, UCL, Dorking, Surrey, United Kingdom
[3]Gagarin Cosmonaut Training Centre, Star city, Russian Federation

k.baykaner@ucl.ac.uk, m.huckvale@ucl.ac.uk, i.whiteley@ucl.ac.uk

Abstract. Automatic systems for estimating operator fatigue have application in safety-critical environments. We develop and evaluate a system to detect fatigue from speech recordings collected from speakers kept awake over a 60-hour period. A binary classification system (fatigued/not-fatigued) based on time spent awake showed good discrimination, with 80% unweighted accuracy using raw features, and 90% with speaker-normalised features. We describe the data collection, feature analysis, machine learning and cross-validation used in the study. Results are promising for real-world applications in domains such as aerospace, transportation and mining where operators are in regular verbal communication as part of their normal working activities.

Keywords: fatigue, speech, computational paralinguistics

1 Introduction

There are a variety of safety-critical environments for which operator fatigue is a significant risk factor, including aerospace, transportation and mining. In response to this risk, a variety of systems have been developed to detect or estimate fatigue. Some of these are accurate but are based on measurements which require expensive or intrusive equipment. However, in some safety-critical environments operators are engaged in regular or constant verbal communication, and in such environments a fatigue monitoring system based on analyzing speech might provide a cheaper and less-intrusive solution.

Existing models for estimating fatigue from speech have tended to focus on predictions of subjective ratings of sleepiness given by the speakers themselves. This paper describes a corpus of speech data from subjects kept awake over a three day period in which they became demonstrably fatigued, and the model training procedures used to classify fatigue based on the objective property of time awake.

2 Background

In safety critical-environments fatigue is a significant risk factor. One of the clearest examples of this is in transportation, where driver fatigue is widely considered to be an important contributory factor in fatal and serious accidents [1], [2]. It is difficult to pinpoint the exact proportion of accidents caused by fatigue, but the consensus of scientists studying safety and accident prevention is that fatigue is the largest identifiable and preventable cause of accidents in transportation, accounting for around 15-20% of all accidents [3]. Fatigue is just as significant a risk in other safety critical settings where vigilance is important, such as aerospace [4].

Models have been developed in the domain of computer-vision to predict and monitor fatigue based on video recordings of operators, and these can be quite accurate (see [5] for a review). Even more accurate ways of monitoring fatigue are possible using intrusive physiological measurements (see [6] for a review of the capacity of electroencephalography, electrocardiography, elektro-okulogram, and pulse oximetry measurements to assess fatigue).

While the vision-based and physiological approaches may be accurate, measuring these features presents a significant challenge to user acceptance in many practical applications because additional, expensive or intrusive equipment is required. By contrast, a cheap and non-intrusive fatigue monitoring system could be implemented if it were possible to predict fatigue by analyzing the voice. This would be particularly useful in those situations requiring drivers or operators to regularly communicate by speaking, (e.g. in aviation, spaceflight, or mining transportation industries).

Existing research has identified vocal correlates with measures of fatigue. For example in [7] it was demonstrated that when subjects were kept awake for a period of 24 hours, the duration of their pauses gradually increased for read speech, and the variation in the 4th formant decreased for sustained vowel sounds.

A variety of models were generated and tested for the Interspeech 2011 speaker state challenge [8] aimed at predicting subjective ratings of sleepiness. For the training and testing data the Sleepy Language Corpus (SLC) was developed, consisting of a mixture of isolated vowels, sustained vowels, commands, and natural speech for 99 speakers. For measures of fatigue, the Karolinska Sleepiness Scale (KSS) was used, which is a subjective scale ranging from 1 (extremely alert) to 10 (extremely sleepy, cannot stay awake). The data was divided into two sets for classification, with the non-sleepy group being all ratings from 1 to 7 (sleepy, but no effort to stay awake), and the sleepy group being all ratings from 8 (sleepy, some effort to stay awake) to 10. The optimal proposed baseline model was able to achieve an unweighted accuracy (UA) of 70.3%, and the winner of the challenge achieved an UA of 71.7% [9]. A higher UA of 82.8% has also been reported in another study based on a subset of the SLC [10].

Since there is no easily accessible ground truth for fatigue it is sometimes unclear what ought to be predicted. The KSS, used in the Interspeech 2011 speaker state challenge, has been validated against performance and EEG measures [11], and although significant correlations were found for most of the measures the highest correlation found (that between KSS and reaction time) had only r=0.57 (standard deviation =

0.25). The authors point out that subjective sleepiness cannot be regarded as a substitute for performance measurements, and similarly that one performance measure cannot usually be substituted for another. Further evidence of the imperfect relationship between subjective scores and performance can be seen in other studies where the correlation between reaction time and KSS scores has been moderate but highly variable (r=0.49-0.71 depending on the subject) [12], or non-existent [13].

In this work models are produced aiming to predict whether or not a subject is fatigued based on sleep latency (i.e. the time the subjects have been kept awake) rather than subjective ratings of sleepiness.

3 Corpus collection and labelling

The goal of corpus collection was to collect speech samples from speakers over an extended period in which they became demonstrably fatigued. The subjects were seven native Russian speakers (six male, one female) who were taking part in a psychological study of the effects of isolation and sleep deprivation. In this study the subjects were isolated and asked to keep awake for over 60 hours. All subjects began the study at 10am on day one and finished the study at 9pm on day three. They were given a range of tasks to occupy their time, including physiological and psychological tests. They were continuously monitored from outside the isolation chamber, but could not communicate with the experimenters.

Speech was collected from the subjects at regular intervals of approximately 6hours. The subjects were asked to read prose from a computer monitor into a Roland R-05 digital recorder sitting on the desk in front of them. The selected prose came from a Russian novel, and was chosen to create a simple and unstressful task for the subjects. The subjects were able to decide themselves how much to read, so recording durations varied between 105 and 495 seconds. Recordings were collected at 24-bit resolution at 44100 samples/sec.

The recordings were post-processed by editing out any speech that was not part of the reading task, by normalizing the signal level and by conversion to 16-bit PCM audio files. In total, 74 speech recordings were made, labelled by speaker and sleep latency (i.e. time since the start of the experiment).

4 Feature extraction

Previous work on predicting sleepiness from speech [e.g. 8] has demonstrated how high-dimensionality feature vectors are useful to capture and represent variation in the signal across conditions. In this work we use a similar approach but using our own feature analysis tools to extract from the whole of each recording a representation of the variation of the speech signal in the frequency domain, the time domain and the modulation domain. The recordings were analyzed and summarized to produce fixed-length feature vectors as follows:

1. The waveform is pre-emphasized and divided into 50ms Hamming-windowed sections overlapping by 10ms.
2. An FFT is applied to each window and a bank of triangular filters is used to calculate a smoothed spectrum on a non-linear frequency scale. The filters are 200mel wide and spaced by 100mel.
3. A cosine-transform of the log-compressed smoothed spectrum is taken to generate 19 MFCC parameters per frame.
4. The first and second temporal differences of the MFCC parameters are computed.
5. The autocorrelation of each window is also computed and interpolated onto a log delay axis.
6. A cosine transform of the log delay autocorrelation function is taken to generate 19 autocorrelation shape parameters per frame.
7. The first and second temporal differences of the autocorrelation shape parameters are computed.
8. The energy of each window is calculated, and the first and second temporal difference is computed.
9. The distributions of the MFCC, autocorrelation and energy parameters are collected over the whole file.
10. The distribution of each parameter was then summarized using the quantile values at 5%, 10%, 25%, 50%, 75%, 90% and 95% together with robust measures of skewness and kurtosis.
11. The audio file was then band-pass filtered between 300 and 3500 Hz, rectified and low-pass filtered at 80Hz to generate a temporal envelope trace. The modulation spectrum of the temporal envelope was calculated using 40 band-pass filters logarithmically-spaced between 0.1 and 50Hz. These parameters were added to the summary statistics parameters generated from the MFCC, autocorrelation and energy analysis.

Ultimately, each file was described by a feature vector containing 1093 parameters.

5 Model construction procedure

In our first study, all recordings prior to 10am on day two were labelled as non-fatigued and the remaining sessions were labelled as fatigued for binary classification. Setting any particular threshold for classification is arbitrary; but setting a threshold of 10am has two distinct benefits. Firstly, it seems reasonable to suggest that any subject would be fatigued after a full 24 hours of wakefulness. Secondly, selecting this threshold results in a corpus with 31 non-fatigued cases and 43 fatigued cases; giving roughly balanced classes with 41.9% and 58.1% of the corpus in each class. Having well balanced classes is important for training a classifier because large imbalances tend to result in models which preferentially predict the majority class.

As in [10], the relative data sparsity makes a speaker-dependent multiple hold-out cross validation approach most appropriate. Specifically a ‘leave one sample out’ cross validation procedure was implemented, where in each iteration a model was trained on data from all subjects with a single sample withheld for validation. The final classification error is calculated by averaging over all 74 classifiers.

For each classifier a support vector machine (SVM) model using a linear kernel was trained, with a margin constraint of 1. Ideally, with a larger corpus available, the approach discussed in [15] would be utilized, and an isolated development set would be used to train SVMs with radial basis function (RBF) kernels, identifying the optimal margin constraint and sigma parameter. With the relatively small corpus, however, it was considered fairer to only use the linear kernel SVM with a fixed margin constraint. Models were trained using the Weka machine-learning toolkit (http://www.cs.waikato.ac.nz/ml/weka/).

Table 1 shows the model performance alongside the performance of two simpler models for comparison: the ZeroR and the OneR [16]. The ZeroR model ignores any features and simply predicts the more common label for every sample. In this case the ZeroR model predicts every sample to be fatigued, and therefore will always have an unweighted accuracy (UA) of 50%. The OneR model utilizes 1-level decision trees based on the single best feature and offers a useful point of comparison since the use of relatively complex machine learning approaches should be justified by showing an improvement over simpler approaches. The performance of the speech model can therefore be considered with reference to the improvement in UA over these simpler models.

Table 1. Comparison of the SVM, ZeroR, and OneR fatigue prediction models using speech features on the 24-hour fatigue task. Positive corresponds to the fatigued class and negative corresponding to the non-fatigued class.

Measure	SVM	ZeroR	OneR
True Positive	36	43	26
False Positive	6	31	12
True Negative	25	0	19
False Negative	7	0	17
Precision	85.7%	58.1%	68.4%
Recall	83.7%	100%	60.4%
Unweighted Accuracy	**82.2%**	50.0%	60.8%

The results show that in this case the linear kernel SVM performed well with a substantially higher unweighted accuracy than either of the simpler models.

6 Using Gaussianized features

Any feature normalisation applied by the SVM models trained in section 5 is performed over all recordings despite the fact that they came from different speakers. Improvements in performance should be possible by explicit normalisation of feature distributions for each speaker prior to training. The goals of feature normalisation are to remove speaker-specific differences in terms of mean, range and distribution shape.

A normalization process called “Gaussianization” [18] was used to transform the speech feature values in such a way that ensures a normal distribution of each feature for each individual speaker. This process maps the empirical feature distributions to a normal distribution by mapping empirical percentiles to values drawn from the inverse cumulative normal distribution.

The use of speaker-specific normalisation has important consequences for practical applications of this fatigue-detection system. Any implementation would now require an enrolment process for each new speaker to establish their personal distribution of feature values. Although the models themselves would be trained from multiple speakers, this limitation effectively renders such systems speaker dependent.

Table 2 shows the results for the model which utilizes Gaussianized speech features. The equivalent performance measures are also shown for the OneR model using the same feature set.

Table 2. Prediction performance on 24-hour fatigue task using models trained with gaussianized features.

Measure	SVM	OneR
True Positive	38	34
False Positive	5	15
True Negative	26	16
False Negative	5	9
Precision	88.4%	69.4%
Recall	88.4%	79.1%
Unweighted Accuracy	**86.1%**	65.3%

The results show an improved UA of 86.1% when using Gaussianized features for the SVM models, compared with 82.2% on raw features. This resulted from increases in both the true positive rate and true negative rate. The simpler OneR model also showed a small increase in UA mostly from an increase in true positives and a decrease in false negatives.

7 Shifting the classification threshold

The choice of 10am as the classification threshold between fatigued and non-fatigued was, as previously noted, an arbitrary one. Here we investigate how well the classification system would operate if the threshold were set earlier to 2am of day two, based on a wakefulness of 16 hours.

Redrawing the classes in this way produces a non-fatigued class size of 24, and a fatigued glass size of 50. This is a more serious imbalance than was the case for the 10am boundary and needed to be addressed. It was decided to use the same strategy as in [9] and utilize the Synthetic Minority Over-sampling Technique (SMOTE) algorithm [14] for generating additional synthetic non-fatigued samples. Using the SMOTE algorithm encourages the classifier training to produce a more generalizable model because the synthetic samples have feature values randomly generated between nearest neighbors; this is particularly important for relatively small corpora such as that considered here. After applying SMOTE, there were 48 non-fatigued samples, giving 98 samples in total and a more balanced corpus with the non-fatigued class accounting for 49.0% of samples.

Table 3 shows the results of carrying out a 98 fold leave-one-out cross validation scheme using the modified boundary.

Table 3. Cross validation fatigue prediction performance with a classification threshold set to 16 hours awake

Measure	SVM Raw	OneR Raw	SVM Gauss	OneR Gauss
True Positive	42	38	46	39
False Positive	9	18	2	14
True Negative	39	30	46	34
False Negative	8	12	4	11
Precision	82.3%	67.9%	95.8%	73.6%
Recall	84.0%	76.0%	92.0%	78.0%
Unweighted Accuracy	**82.6%**	69.3%	**93.9%**	74.4%

The cross-validation performance of the SVM models using the 2am threshold is similar to those obtained using the 10am threshold, with 82.6% and 82.2% unweighted accuracy achieved respectively when using raw features, and 93.9% and 86.1% when using gaussianized features. The performance cannot be directly compared since the former model was trained without synthetic data being added. In both cases the simpler OneR models performed more poorly.

8 Split validation

Although good performance was shown by the models in the previous sections, there is still uncertainty about whether the choice of leave-one-out cross-validation is giving unfair advantage to the classifier. This is particularly relevant for SVM models which operate by selecting and retaining specific training vectors. It may be that good performance came from retaining vectors which happen to be close to each test vector.

For more robust assessment we would ideally use a validation set of new speakers independent from those used in training, but our small corpus size makes this difficult. Instead we have implemented a split validation procedure to give an indication of the expected accuracy when making predictions on new data. If performance after split-validation was significantly worse than leave-one out cross-validation then we would need to be concerned about the robustness of our approach.

The split validation test was first performed on the 10am threshold data set. The data were split into a 2/3 training (49 samples) and 1/3 validation (25) set with samples selected at random. A linear kernel SVM and a OneR model was generated based on the training data for the raw and Gaussianized feature sets, and performance was measured by testing predictions on the validation set. Table 4 shows the performance for these models.

Table 4. Prediction performance on the 24hr fatigue task using separated training and validation data sets.

Model	Precision	Recall	Unweighted Accuracy
Raw features			
ZeroR	64.0%	100.0%	50.0%
OneR	72.7%	50.0%	58.3%
SVM	86.7%	81.3%	**79.5%**
Gaussianized features			
OneR	70.0%	43.8%	55.2%
SVM	100%	87.5%	**93.8%**

The split validation performance indicated that the SVM linear kernel model based on gaussianized features produced the highest UA of 93.8% (cf. 86.1% for leave-one-out), with a lower UA of 79.5% (cf. 82.2% for leave-one-out) achieved using raw features.

The split validation procedure was then repeated on the 2am threshold corpus described in section 7. Table 5 shows the performance.

Table 5. Prediction performance on the 16hr fatigue task using separated training and validation data sets.

Model	Precision	Recall	Unweighted Accuracy
Raw features			
ZeroR	56.0%	100.0%	50.0%
OneR	54.5%	42.9%	44.2%
SVM	85.7%	85.7%	**83.8%**
Gaussianized features			
OneR	62.5%	71.4%	58.4%
SVM	92.3%	85.7%	**88.3%**

Using raw features, the UA was 83.8% (cf. 82.6% for leave-one-out). Using Gaussianized features, the UA was 88.3% (cf. 93.9% for leave-one-out).

Generally model performance held-up well under split-validation, with performance varying from 7.7% better to 5.1% worse compared to cross-validation. As with the cross-validation, the split validation performance of the models with the 2am threshold were similar to those obtained using the 10am threshold.

It is likely that the variation in unweighted accuracy from the split validation procedure observed here is a result of the relatively small corpus size, which in turn produces a higher variability in performance measures. In general, this suggests that performance differences on the order of 8% are possible when applying this modelling approach to new data.

9 Conclusion

Speech recordings were gathered from subjects undergoing three days without sleep. Over this period the subjects became increasingly fatigued. Features were generated by analyzing speech recordings collected at regular intervals and these features were used to train SVM classifiers predicting whether the subject was fatigued or non-fatigued based on sleep latency. The model training results show that the use of Gaussianized features significantly improves prediction accuracy; but it should be noted that practical implementations using Gaussianized features require a speaker enrolment stage for each new speaker. On separated validation data the unweighted accuracy of the SVM classifiers was around 80% for raw features and 90% for Gaussianized features. Good performance was obtained for fatigue thresholds set at either 16hours or 24hours of sleep deprivation.

It may be inconvenient in some applications to require several minutes of speech in order to predict fatigue. However, the generated features can be produced based on much shorter recordings, and similar results have been obtained utilizing speech excerpts of 20 seconds. Further work should aim to determine whether tasks based on spontaneous speech would be more revealing of fatigue compared to the read speech used here.

This work shows that using features generated from recordings of speech, meaningful predictions can be made of sleep latency as an objective measure of fatigue. This is promising for the development of decision-making aids applied in safety-critical environments where the fatigue level of an operator is an important risk factor.

10 Acknowledgements

The authors would like to acknowledge the European Space Agency and University College of London who are jointly responsible for funding this work.

11 References

1. Dobbie, K.: Fatigue-related crashes: an analysis of fatigue-related crashes on Australian roads using an operational definition of fatigue. Australian transport safety bureau (OR23), (2002).
2. FMCSA: Regulatory impact analysis – hours of service final rule. Federal motor carrier safety administration, December, 2011.

3. Åkerstedt, T.: Consensus statement: Fatigue and accidents in transport operations. Journal of sleep research, 9, 395-395 (2000).
4. Rosekind, M., Gander, P., Miller, D., Gregory, K., Smith, R., Weldon, K., Co, E., McNally, K., Lebacqz, J.: Fatigue in operational settings: examples from the aviation environment. J. Human Factors. 36, 327-338 (1994).
5. Barr, L., Howarth, H., Popkin, S., Carroll, R.: A review and evaluation of emerging driver fatigue detection measures and technologies. John A. Volpe National Transportation Systems Center (2005).
6. Begum, S.: Intelligent driver monitoring systems based on physiological sensor signals: a review. IEEE annual conference on intelligent transportation systems (ITSC) (2013).
7. Vogel, A., Fletcher, J., Maruff, P.: Acoustic analysis of the effects of sustained wakefulness on speech. Journal of the acoustical society of America 128, 3747-3756 (2010)
8. Schuller, B., Batliner, A., Steidl, S., Schiel, F., Zrajewski, F.: The Interspeech 2011 Speaker state challenge. Proceedings of Interspeech 2011, 2301-2304 (2011)
9. Huang, D., Ge, S., Zhang, Z.: Speaker State Classification Based on Fusion of Asymmetric SIMPLS and Support Vector Machines. Proc Interspeech 2011, 3301-3304 (2011).
10. Krajewski, J., Batliner, A., Golz, M.: Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. Behavior Research Methods, 41, 795-804 (2009).
11. Kaida, K., Takahashi, M., Akerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., Fukasawa, K.: Validation of the Karolinska sleepiness scale against performance and EEG variables. Clinical Neurophysiology. 117, 1574-81 (2006).
12. Gillberg, M., Kecklund, G., Akerstedt, T.: Relations between performance and subjective ratings of sleepiness during a night awake. J. Sleep, 17, 236-241 (1994).
13. Åhsberg, G., Kecklund, G., Åkerstedt, T., Gamberale, F.: Shiftwork and different dimensions of fatigue. International Journal of Industrial Ergonomics. 26, 457-465 (2000).
14. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research. 16, 321-357 (2002).
15. Hsu, C., Chang, C., Lin, C.: A practical guide to support vector classification. Department of computer science technical report, National Taiwan University (2010).
16. Holte, R.: Very simple classification rules perform well. Journal of Machine Learning. 11, 63-91 (1993).
17. Williamson, A., Lombardi, D., Folkard, S., Stutts, J., Courtney, T., Conner, J.: The link between fatigue and safety. Accident Analysis & Prevention. 43, 498-515 (2011).
18. Chen, S., Gopinath, R.: Gaussianization. Proc. NIPS 2000, Denver Colorado (2000).