*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Prediction of Sleepiness Ratings from Voice by Man and Machine

*Mark Huckvale, András Beke, Mirei Ikushima*

Speech, Hearing and Phonetic Sciences, University College London, U.K.
m.huckvale@ucl.ac.uk

## Abstract

This paper looks in more detail at the Interspeech 2019 computational paralinguistics challenge on the prediction of sleepiness ratings from speech. In this challenge, teams were asked to train a regression model to predict sleepiness from samples of the Düsseldorf Sleepy Language Corpus (DSLC). This challenge was notable because the performance of all entrants was uniformly poor, with even the winning system only achieving a correlation of r=0.37. We look at whether the task itself is achievable, and whether the corpus is suited to training a machine learning system for the task. We perform a listening experiment using samples from the corpus and show that a group of human listeners can achieve a correlation of r=0.7 on this task, although this is mainly by classifying the recordings into one of three sleepiness groups. We show that the corpus, because of its construction, confounds variation with sleepiness and variation with speaker identity, and this was the reason that machine learning systems failed to perform well. We conclude that sleepiness rating prediction from voice is not an impossible task, but that good performance requires more information about sleepy speech and its variability across listeners than is available in the DSLC corpus.

**Index Terms**: Sleepiness, Voice, Machine Learning, Paralinguistics

## 1. Introduction

The 2019 Computational Paralinguistics challenge [1] included a continuous sleepiness rating prediction task based on the Düsseldorf Sleepy Language Corpus (DSLC). In this challenge teams were asked to build a machine learning system to predict self-ratings of sleepiness of speakers from short audio excerpts of their speech. The ratings were on a scale of 1-9 using the Karolinska Sleepiness scale (KSS) [2] varying between "extremely alert" to "very sleepy". Performance of the systems reported in the challenge were very poor, with the winning system only achieving a correlation of r=0.37 with the human ratings. Table 1 summarises some of the different systems and performance figures.

The poor performance of machine learning systems at this task demands explanation. Is it really the case that sleepiness ratings cannot be predicted accurately from speech? Are the feature representations of the signals used by the systems inadequate in some way? Is this a task for which machine learning is not suited? Are there problems with the audio or the labelling of the corpus?

The outcome of the 2019 sleepiness challenge seems particularly poor when compared to the 2011 sleepiness challenge [3]. In 2011 the challenge was to classify speech recordings into sleepy vs non-sleepy, using KSS ratings of 7 and below as non-sleepy, and 8 and above as sleepy. In that challenge the baseline systems achieved an accuracy of about 70%, while a later study using the same speech data and task achieved a classification accuracy of over 80% [4]. Could the failure of the 2019 challenge be because of the switch from a classification task to a regression task?

In this paper we investigate the DSLC corpus in more detail. Our goals are to understand the reasons behind the poor performance of machine learning systems for predicting sleepiness ratings from speech using these data. The outcomes should be useful in building better systems for assessing sleepiness which may be useful in detecting the fatigue of operators in safety-critical jobs.

In section 2 we describe the corpus and how the reference ratings were obtained. In section 3 we present the results of a new perceptual experiment in which we ask a panel of listeners to provide sleepiness ratings for 90 recordings from the corpus constituting a new test set. We look at the performance of human listeners on labelling sleepiness, and discuss what that tells us about difficulty of the task. In section 4 we perform a number of statistical analyses of the corpus to understand why machine learning methods trained on the corpus performed so poorly. In section 5 we discuss the implications of the findings for future work in sleepiness rating prediction.

Table 1.*Published performance figures for machine prediction of sleepiness ratings in the Interspeech 2019 challenge*

| **System** | **Correlation on Development Set** |
|---|---|
| OpenSMILE + SVR (Baseline) [1] | 0.251 |
| Bag of audio Words + SVR (baseline) [1] | 0.269 |
| Deep Learning autoencoder + SVR (baseline) [1] | 0.261 |
| OpenSMILE + SVR [5] | 0.327 |
| Fisher Vectors + SVR [5] | 0.355 |
| Bag of audio words + SVR [5] | 0.300 |
| OpenSMILE+Fisher Vector fused [5] | 0.367 |
| OpenSMILE+BoAW+Fisher Vector fused [5] | **0.368** |

## 2. Sleepy Language Corpus

The Düsseldorf Sleepy Language Corpus was recorded at the Institute of Experimental Psychophysiology, Düsseldorf, and the Institute of Safety Technology, University of Wuppertal, Germany. The corpus used in the Interspeech 2019 challenge [1] included recordings from 915 German speakers (364 females, 551 males, age from 12 to 84 years, mean age 27.6 ±



http://dx.doi.org/10.21437/Interspeech.2020-1601

11.0 years). The recordings of different reading passages and spontaneous speaking tasks were made in quiet rooms with the tasks presented on a computer in front of the participants. Audio files were recorded at 44100 samples/sec and down-sampled to 16000 samples/sec, with 16-bit quantisation. A session of one subject lasted from 15 minutes to 1 hour and recordings took place at different times of day between 6am and midnight. Each participant had to rate their sleepiness on the KSS, and ratings were also made by two expert raters. The scores from self-assessment and observers were averaged to form the reference sleepiness values. It is not stated in the available documentation exactly how the averaging was performed, but in a previous study by the creators of the corpus [6], the expert raters were present during the recording and made their ratings contemporaneously, quote:

> "A well established, standardised subjective sleepiness questionnaire, the Karolinska Sleepiness Scale (KSS), was used by the subjects (self-assessment) and by the three assistants who had supervised the experiments, using all available information (audio/video/context); they had been formally trained to apply a standardised set of judging criteria." [6]

In total there were 16,462 recordings from the 915 speakers, but in this investigation we only have access to labels for the training subset (5,546 recordings) and the development subset (5,328 recordings).

## 3. Perceptual Experiment

Ninety-nine recordings from the development subset of the corpus were used in a perceptual experiment. Recordings were evenly selected across the different ratings, with 1 recording of each rating level used for training listeners and 10 recordings of each rating level used for test. 26 English-speaking listeners took part using a web interface, see Fig.1. Each listener rated each of the files on a scale of 1 to 9 using the Karolinska sleepiness rating scale rubric.

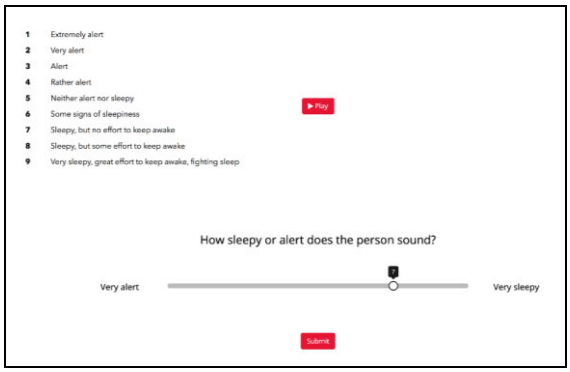


Figure 1. *Listening Test Screenshot*

The raw ratings for each listener were first normalised to zero mean and unit variance to remove some variability across listeners in the way in which they used the scale. The normalised responses for all listeners for all recordings are shown in Fig.2. The violin plot shows the distribution of normalised scores for each rating level used in the test data. Looking at raw scores alone, there is much variation in listener score for every rating level, with the overall correlation being only r=0.249. In terms of inter-rater agreement, Kendall's coefficient of concordance is only 0.112, indicating considerable diversity of opinion among the raters.

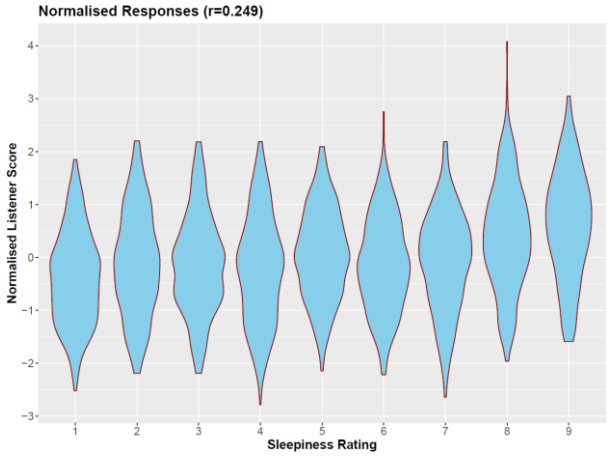


Figure 2. *Normalised listener ratings of speech files*

We can use the "wisdom of the crowd" to refine the listener ratings by averaging normalised ratings across all 26 listeners. Fig.3 shows the mean normalised listener responses for each recording as a function of rating label. Each dot on this plot represents one recording. The boxes show the entire range, the inter-quartile range and the median value of the score distribution. The averaged listener ratings show a much higher correlation with the supplied ratings than individuals, with r=0.72. Variation of the labelling was assessed using Friedman's rank sum test, which shows significant variation in responses across ratings ($\chi 2$=114, df=8, p <0.001). However, a post-hoc analysis of the Friedman test shows that there were only three significantly different rating groups: 1 | 2,3,4,5,6,7 | 8,9. This is visually apparent in Fig.3.

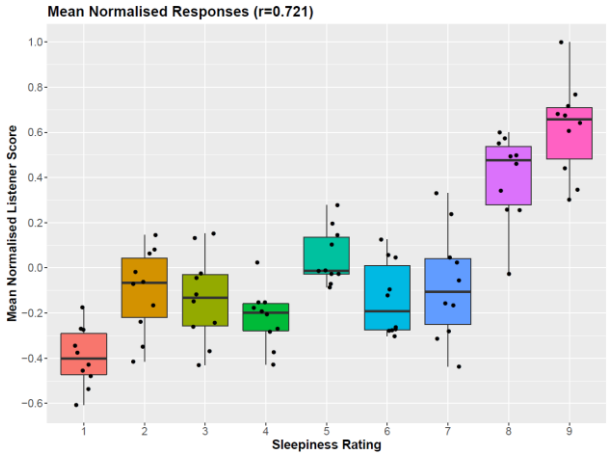


Figure 3. *Mean normalised listener ratings per recording*

What do we learn from this perceptual experiment? Firstly, this is not an easy task for untrained listeners. Inter-rater reliability was low, and only the average score had reasonable correlation with the supplied ratings. This was despite the fact that the self-ratings had been smoothed by averaging with the perceived judgments of expert listeners, which might have been expected to bias the ratings towards the auditory characteristics of the speech. Secondly, there is no support for the hypothesis that listeners can use the 1-9 rating scale effectively. The data seem to show that all listeners were doing was dividing the recordings into a three-way groupings

of aroused/normal/sleepy. Interestingly, the most sleepy group combines ratings 8 & 9 which matches the threshold used in the 2011 sleepiness challenge [3], where ratings 7 and below were classed non-sleepy, and ratings 8 and above were classed sleepy.

We can use the listener responses to simulate a classification task into sleepy vs. non-sleepy. Using a threshold of 7.5 for the ratings and a threshold of 0.26 for the mean listener response, we obtain the classifications shown in Table 2. While performance here is extremely high, it must be remembered that the corpus ratings may themselves be biased towards human listener perceptions.

Table 2 *Post-hoc sleepy/non-sleepy classification using listener responses*

| | **Non-sleepy** | **Sleepy** |
|---|---|---|
| **Non-sleepy** | 68 | 2 |
| **Sleepy** | 2 | 18 |
| Accuracy=95.6% UAR=93.6% | | |

In this section we have explored the behaviour of human listeners on the sleepiness rating task on the test set. In the next section we will explore the characteristics of the whole corpus to identify the causes of the poor machine learning performance.

## 4. Corpus Analysis

In the light of the good performance of human listeners at predicting sleepiness, the poor performance of machine learning systems could be due to a number of different factors: (i) differences in the way in which sleepiness affects the voice across speakers, (ii) differences in the make-up of the different partitions of the corpus, so that what is learned from one is inapplicable to the other, (iii) other interactions between speakers, ratings and corpus partitions.

Speaker labels are not available for the corpus, yet we know that there are only 915 speakers in the 16462 recordings, so multiple recordings must have been made of each speaker. It is also reasonable to assume that the speakers in the three corpus partitions sets are disjoint, which means that there are about 300 speakers and about 18 recordings/speaker on average in each partition.

The first task is to check the degree of overlap between the train and development partitions in terms of speakers. We do this in an informal way at first by projecting feature analyses of the recordings into 2 dimensions and colouring them by whether they are in the training or development set.

For data representation we use x-vectors [7]. X-vectors are deep neural network embeddings learned from a speaker recognition task. They represent the state of a hidden layer in a neural network designed to predict a probability distribution over a set of speakers from an audio feature representation. In this task we use the x-vector system available as part of the Kaldi toolkit, which was trained on augmented Switchboard, Mixer 6, and NIST SRE corpora (https://kaldi-asr.org/models/m3). This delivers 512 dimensional vectors from each audio recording.

Figure 4 shows the projection of the x-vectors into 2 dimensions using the tSNE method [8], using different colours for the training and development partitions. The tSNE method aims to preserve in the lower dimensional space the local clustering found in the original high dimensional space. What can be clearly seen in the figure is that there are many small clusters which are in the main all from the training set or all from the development set. There are few clusters which seem to contain both red and blue dots. We conclude that, as expected, the speakers used in the training data and in the development data partitions are different.

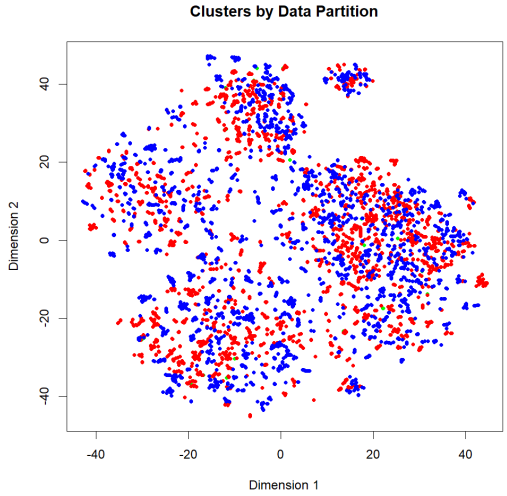


Figure 4. *2D projection of x-vector features, red=training set, blue=development set*

Figure 5 shows a similar analysis, but this time coloured according to sleepiness rating. What is interesting here is that the clusters - which we believe represent different speakers - only seem to present a small range of sleepiness ratings (small range of colours). That is, we conclude that the ratings provided by each speaker did not vary much across the multiple recordings found in the corpus.

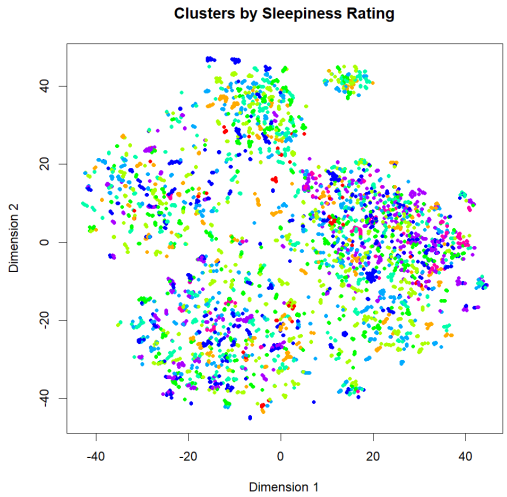


Figure 5. *2D projection of x-vector features coloured by sleepiness ratings*

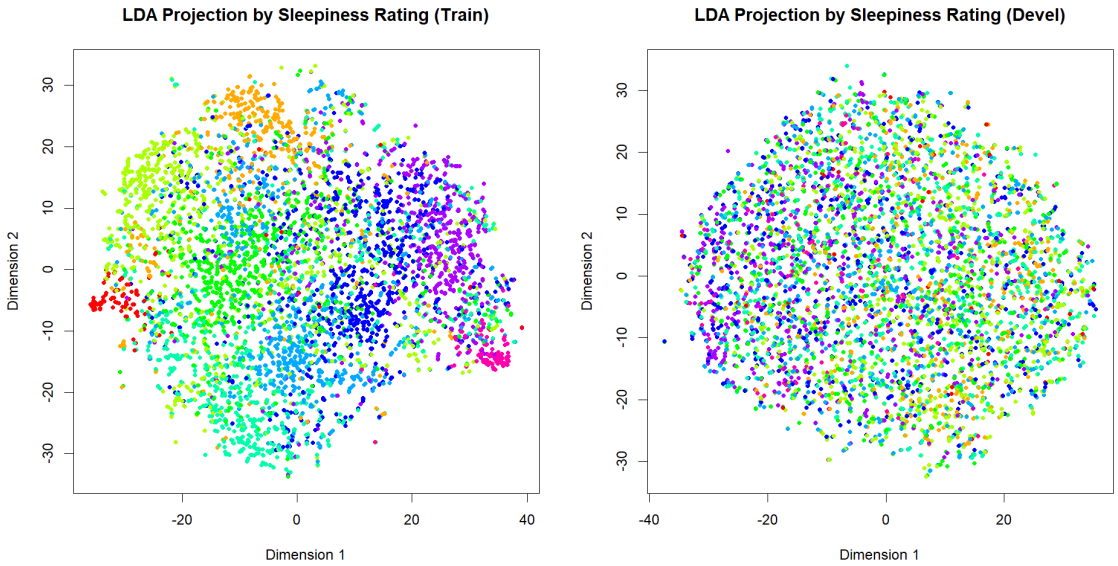


Figure 6. *2D tSNE projections of LDA projections of x -vector features, coloured by sleepiness rating*

Next, we look at how sleepiness ratings are distributed through the x-vector space in the training and development sets separately. To do this we perform a linear discriminant analysis (LDA) on the x-vectors in the training partition using the ratings as labels. Then using the identified directions of the discriminants we project the training data and the development data to two dimensions using tSNE. Figure 6 shows the projections for the two partitions. What is very clear from the figure is that the LDA discriminants do a good job of separating out the different ratings groups for the training data (from which they were calculated), but those same discriminants are much worse at separating the ratings groups found in the development data.

In this section we have shown important differences between the training and development partitions in the sleepy-language corpus. It appears that, as expected, there are different speakers in the two partitions, but in combination with that, there is little variation in sleepiness ratings given by each speaker. This means that sleepiness and speaker identity are fundamentally linked in the corpus, and that any simple approach to analysis of sleepiness in the training partition will inevitably turn into a speaker recognition system. Such a system will then perform poorly on a development partition containing different speakers.

## 5. Conclusions

This goals of this study were to find an explanation for the poor performance of machine learning systems for predicting sleepiness ratings from speech in the DSLC, and to suggest directions for future work.

Through the analysis of the corpus itself in section 4, we have seen that a major problem is the confounding of speaker identity and sleepiness ratings in the corpus. Each corpus partition contains different speakers, and each speaker only produced a narrow range of sleepiness ratings. This makes it very hard to learn features of sleepiness from the training set without at the same time learning features of identity. When those features are exploited by the prediction model, they may work well to measure similarity between speakers in the test set to speakers in the training set, but it is not necessarily the case that those similar speakers have similar sleepiness ratings.

While we might conclude from this analysis that the corpus itself is fundamentally compromised for machine learning, the result of the human listening experiment described in section 3 is much more encouraging. The listeners as a group were able to separate the recordings in the test set into three groups on the basis of sleepiness. When applied to the binary task of distinguishing sleepy from non-sleepy, accuracy was over 90%. It is clear that the listeners as a group had access to knowledge that helped them solve this problem without needing to learn from the training set. That knowledge might be in two forms: knowledge about how sleepiness changes the way in which speakers speak, and knowledge about how speech varies across individuals. Human listeners are likely to be highly attuned to changes in the voice relevant for the observation of speaker state. But not only this, they are able to exploit that knowledge without previous exposure to the speaker. The machine learning systems for predicting sleepiness that were submitted to the challenge were limited by the fact that the only knowledge they had of variability within and across speakers came from the DSLC training set – and that variation in sleepiness and identity were confounded in this corpus.

Future work in this area will benefit greatly from data analysis methods that separate out characteristics of identity and sleepiness in voice. This type of analysis could be based on the kind of factor analysis used in speaker recognition to separate out speaker identity from speaking environment. However it will require speech corpora that are more varied within speakers; and with labelling for speaker as well as labelling for sleepiness.

## 6. Acknowledgements

The authors are grateful to the creators of the Düsseldorf Sleepy Language Corpus and the organisers of the 2019 Interspeech Paralinguistics Challenge that have made this study possible.

# 7. References


[1] Björn W. Schuller, Anton Batliner, Christian Bergler, Florian B. Pokorny, Jarek Krajewski, Margaret Cychosz, Ralf Vollmann, Sonja-Dana Roelen, Sebastian Schnieder, Elika Bergelson, Alejandrina Cristià, Amanda Seidl, Anne Warlaumont, Lisa Yankowitz, Elmar Nöth, Shahin Amiriparian, Simone Hantke, Maximilian Schmitt: The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. *Proceedings INTERSPEECH 2019, ISCA, Graz, Austria, 2019.*

[2] T. Åkerstedt, M. Gillberg, Subjective and objective sleepiness in the active individual. *Int J Neurosci, 52 (1990), pp. 29-37*

[3] Bjorn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, Jarek Krajewski, The INTERSPEECH 2011 Speaker State Challenge. *Proceedings of Interspeech 2011, Florence, Italy.*

[4] Bilge Günsel, Cenk Sezgin, Jarek Krajewski, Sleepiness detection from speech by perceptual features. *Proceedings of ICASSP 2013, Vancouver, Canada.*

[5] Gabor Gosztolya, Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds. *Proceedings of Interspeech 2019, Graz, Austria.*

[6] Florian Honig, Anton Batliner, Elmar Noth, Sebastian Schnieder, Jarek Krajewski, Acoustic-Prosodic Characteristics of Sleepy Speech – Between Performance and Interpretation. *Speech Prosody 7, 2014.*

[7] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, Sanjeev Khudanpur, X-vectors: robust DNN embeddings for speaker recognition. *Proceedings of ICASSP 2018, Calgary, Canada.*

[8] Laurens van der Maaten, Geoffrey Hinton, Visualizing Data using t-SNE. *Journal of Machine Learning Research 9 (2008) 2579-2605.*