Large Sleepy Reading Corpus (LSRC): Applying Read Speech for Detecting Sleepiness

Jarek Krajewski[1,2], Sebastian Schnieder[3], Christopher Monschau[2], Raphael Titt[4], David Sommer[5], Martin Golz[5]

[1] Institute of Safety Technology, University of Wuppertal, 42119 Wuppertal
[2] Engineering Psychology, Rhenish University of Applied Science Cologne, 50825 Cologne
[3] Business Psychology, University of Applied Science for Media, Communication, and Business, 13355 Berlin
[4] Social Psychology, University of Tuebingen, 72074 Tübingen
[5]Applied Computational Intelligence, University of Applied Sciences Schmalkalden, 98574 Schmalkalden

Email: {krajewsk}@uni-wuppertal.de

Abstract

This paper describes a Large Sleepy Reading Corpus (LSRC) based on sleep deprivation data (N=402; total duration 22 h). During the sleep deprivation, a standardized self-report scale was used just before the recordings to determine the sleepiness state. The speech material consisted of different reading passages. In order to investigate sleepiness induced speech changes, a standard set of spectral and prosodic features was extracted from recordings. After applying a standard openSMILE feature set, and a SVM regression we achieved correlation coefficients of .44 for male and .53 for female speaker.

1 Introduction

Sleepiness has been widely accepted as a significant cause in a variety of traffic accidents [1-3] and work contexts (e.g., safety sensitive fields such as chemical factories, nuclear power stations, and air traffic control [4,5]). Measuring sleepiness has been recognized as an important factor for the prevention of a broad range of traffic accidents [1, 2, 6-8]. Hence, many efforts have been reported in the literature for developing real-time sleepiness detection systems. These systems mainly focus on visual information such as (a) instability of pupil size [9], eye blinking [10-12], eyelid movement [7], and saccade eye movement [13, 14] as well as (b) gross body movement, head movement, mannerism, and facial expression in order to characterize a driver's state of alertness [15]. Several disadvantages of those video-based instruments are a lack of robustness against environmental and individual-specific variations (e.g., bright light, wearing correction or sunglasses, occlusions, or anatomic variations such as small palpebral fissures) and a lack of comfort and longevity due to electrode sensor application.

In contrast, voice analysis is non-obtrusive, easy to apply even under several extreme environmental conditions (bright light, darkness, high humidity, high or low temperature, vibration, e.g. [16]), and omnipresent in Human-Computer-Interaction. If the user shows unusual sleepiness states, giving feedback about this fact would make the communication more empathic and human-like. This enhanced naturalism might improve the acceptance of these systems. Furthermore, it may result in better comprehensiveness, if the system output is adapted to the user's actual sleepiness impaired attentional and cognitive resources.

These changes summarized in the cognitive-physiological mediator model of sleepiness-induced speech changes [21] provide the first insight into and theoretical background for the development of acoustic measurements of sleepiness in read speech. Sleepiness-related cognitive-physiological changes such as decreased muscle tension or reduced body temperature can indirectly influence voice characteristics according to stages of speech production [13, 14, 16]. Moreover, the following changes might appear in sleepy speaker: reduced cognitive processing speed (central slowing hypothesis) impaired speech planning and impaired neuromuscular motor coordination processes (psychomotor slowing), impaired fine motor control and slowed articulator movement, slackened articulation and slowed speech. Nevertheless, little empirical research has been done to examine these processes mediating between sleepiness, speech production, and acoustic features.

Previous research mainly focuses on sleepiness detection under high intensities of sleepiness e.g. Interspeech 2011 Speaker State Challenge [17, 21-26]. When applying the results in daily life settings, this strong bimodal distribution of sleepiness states (totally sleepy and completely alert) might lead to an overestimation of accuracy. Thus, enlarging existing databases by a greater diversity of speaker, and more moderate sleepiness scores - which might be additionally more relevant for the early predicting of accidents - is essential. The major drawback of the current state of the art in acoustic sleepiness research is the lack of large data corpora containing several hundreds of speakers. This drawback is addressed in this study by using different read tasks and a large corpus of speakers.

Another open question within sleepy speech research is the adaption of sleepiness models to different speaking styles and speech tasks. Monolog tasks (e.g. picture description task or free recall) involves associative processes and long-term memory – on the other hand - dialogs task require more working memory. Reading aloud longer coherent texts in contrast to short sentences requires much more cognitive resources (e.g. working memory, attention). Thus, read speech differs from free speech in its prosodic and spectral characteristics. For example, mean F0 is significantly lower and F0 range is significantly greater for speaking than tier reading [19]. Considering the cognitive complexity [18] involved in reading longer texts, several additional processes of cognitive-motorical speech production are probably affected by sleepiness. In sum, little is known about the effectiveness of reading long text

passages (e.g. within a fit-for-duty application scenario) for the detection of sleepiness. Thus, the paper is organized as follows Section 2 describes the speech database, section 3 feature extraction, and section 4 regression algorithm employed and results. After providing the results of the sleepiness detection in Section 4, conclusions and future work are discussed in Section 5.

2 Large Sleepy Reading Corpus

We conducted a sleep deprivation study with 402 participants (252 f, 150 m). We used a naturalistic audio file sample of read speech from readers with normal intensities of sleepiness measured by the Karolinska Sleepiness Scale (KSS). The subjects under sleep deprived conditions slept less than 6 hours prior the recording session. Every speaker took part in different reading tasks. After elimination of incomplete audio recordings and corrupted files a total 1072 speech recordings was incorporated in this data corpus (total duration: 22 h of recordings).

The mean age of subjects was 31.0 years, with a standard deviation of 13.1 years and a range of 15–66 years. The audio files were recorded with 44.1 kHz, down-sampled to 16 kHz with a quantization of 16 bit. The speech data consisted of 7 read speech tasks. The reading tasks employed in this database required participants to read:

- ”the North Wind and the Sun”,
- “Rainbow Passage”,
- “Butter Story” (widely used within phonetics, and speech pathology),
- two passages of the novel “Homo Faber”,
- a german newspaper article, and
- a scientific text about phonetics.

For the study, the available recordings were split into two sets (male set and female set), speaker-independently in ascending order of subject ID into training, development, and test instances. This division not only ensures speaker-independent partitions but also provides for stratification by study setup (environment and degree of sleep deprivation). Out of the 1072 utterances (354 m, 718 f), 424 (137 m, 287 f) were assigned to the training set, 318 (103 m, 215 f) to the development set, and 330 (114 m, 216 f) to the test set.

Table 1. *Distribution of long reading passages in LSRC dataset divided by gender and train, development and test set.*

LSRC	All	Train	Develop	Test
Male	354	137	103	114
Female	718	287	215	216
Total	1072	424	318	330

Table 2. *Mean and standard deviation for KSS in the respective sets.*

LSRC		Male	Female
Train	M	5.21	3.99
	SD	1.74	1.39
Develop	M	4.04	4.77
	SD	1.65	1.56
Test	M	3.40	6.01
	SD	1.34	1.46
Train+Devel	M	4.71	4.32
	SD	1.79	1.51
Train+Devel+ Test	M	4.34	4.67
	SD	1.76	1.72

A well established, standardized subjective sleepiness questionnaire, the Karolinska Sleepiness Scale (KSS), was used by the subjects for assessment of subjective sleepiness. The version used in the present study scores range from 1–10: extremely alert (1), very alert (2), alert (3), rather alert (4), neither alert nor sleepy (5), some signs of sleepiness (6), sleepy, but no effort to stay awake (7), sleepy, some effort to stay awake (8), very sleepy, great effort to stay awake, struggling against sleep (9), extremely sleepy, cannot stay awake (10).

3 Feature Extraction

We apply the standard speech features derived from the openSMILE feature extraction toolkit for the detection of sleepiness states [21]. The feature set is built from sets of low-level descriptors and several corresponding sets of functionals applied on the recording level for each LLD set. The LLD sets are given in table 2. The 6551 acoustic-prosodic features containing set can be divided into smaller subsets consisting of homogenous feature groups: Energy-based features, F0-based features, Melspec-based features, F-band-based features (0-250 Hz, 250-650 Hz, 1-4 kHz), MFCC features, spectral features, voice probability-based features, Zero-Crossing-Rate based features.

For transparency and easy reproducibility, we use the WEKA data mining toolkit for regression [20]. As regression algorithm, we chose Support Vector Machines regression (SMOReg) with RBF kernel. For providing a baseline regression no parameter optimization or computerized feature selection was performed. All feature values have been normalized by the SMOReg implementation in WEKA. The results of the SVM regression when training on the train partition of the database and testing on the respective development set is shown in table 4. Also, the regression results for the combined training and development set are reported when tested on the test set. Correlation coefficients are reported as an evaluation metric for each experiment.

Table 3. *Low-level descriptor (LLD) overview.*

energy related LLD
Zero-Crossing Rate
LogEnergy
Melspec
spectral LLD
MFCC 1- 12
Spectral Roll Off 0.25, 0.50, 0.75, 0.90
Spectral flux, centroid, maxPos, minPos
Spectral Energy 0-250 Hz, 0-650 Hz, 250-650 Hz, 1000-4000 Hz
voice related LLD
F0, F0 envelope
voice probability

4 Experiments and Results

4.1 Analysis of single features

As table 4 shows correlations between extracted features and respective KSS-score is higher among the male set, similar to the findings in [22]. Interestingly, Melspec-based features appear to be of relevance for both sets, while some features are unique for the respective set. ZCR-based features not only correlate relatively high for the male set, also they are not represented at all in the top ten features for the female set.

Table 4. *Correlation coefficients between sleepiness and acoustic features divided by gender.*

LSCR	**r**	**feature**
Male	.35	zcr_sma_de_peakMeanMeanDist
	.34	zcr_sma_de_de_peakMeanMeanDist
	.32	zcr_sma_stddev
	.32	zcr_sma_percentile98.0
	.31	melspec_sma_de_de[8]_nzqmean
	.31	melspec_sma_de_de[8]_qmean
	.31	melspec_sma_de_de[8]_variance
	.30	mfcc_sma_de[1]_peakMeanMeanDist
	.30	zcr_sma_percentile95.0
	.30	mfcc_sma[1]_nzqmean
Female	.26	spectralMaxPos_sma_iqr2-3
	.25	melspec_sma_de_de[23]_linregerrQ
	.25	melspec_sma_de_de[23]_variance
	.25	melspec_sma_de[19]_variance
	.24	melspec_sma_de[16]_qmean
	.23	melspec_sma_de[16]_nzqmean
	.23	spectralRollOff75.0_sma_qmean
	.23	spectralRollOff75.0_sma_nzqmean
	.22	spectralRollOff25.0_sma_de_de_iqr1-3
	.22	melspec_sma_de_de[23]_nzqmean

Table 5. *Pearson correlation as evaluation metric of regression performance for the LSRC based sleepiness detection.*

LSRC	**Train vs. Develop**	**Train+Develop vs. Test**
Male	.28	.44
Female	.41	.53

4.2 Regression Experiments

For robust and efficient estimation, Support Vector Machine regression (SMOReg) with RBF kernel is trained. The results are given in table 5. Regression is computed for male and female speakers separately and evaluated regarding their actual performance on unknown test data. Table 2 shows class distribution among the different sets. As one can see, the classes are distributed unevenly for the data sets Train and Develop, but highly uneven in Test set. This uneven distribution is of high difficulty for the regression algorithm.

Correlation coefficients are reported for evaluation purposes. The female speaker set performs with r = .41 on train vs develop set and with r = .53 on train+develop vs Test. The male speaker set performs with r = .28 on train vs develop set and with r = .44 on train+develop vs test. The weaker performance of the male speaker set could be explained by the smaller sample size and the lacking prevalence of sleepy speech samples. This would also explain the relative large rise in performance from development set to test set.

5 Conclusions

From a research-infrastructural-based view, the aim of the study is to build a large sleepy reading corpus (LSRC), containing many hundreds of speakers reading long text passages under mild sleepiness. A further goal of this study was to explore whether gender-specific models can predict sleepy speech. The main findings of the present study may be summarized as follows. First, a standard brute-force feature set is extracted from previously recorded read speech. Then, the resulting features are categorized in two gender-specific sets consisting of the low-level descriptors and their respective statistical functionals. Regression for each speaker set is computed afterward and evaluated towards their respective performance in identifying sleepy speech. In conclusion, the algorithm for female speaker set performs slightly better. By considering the instance distribution in the speaker sets, evidence is given that for training a robust prediction algorithm large sample sizes and even class distributions are mandatory. Nevertheless, the conducted experiment also give evidence that it is possible to classify sleepiness from read speech in a subject-independent design.

However, there are some limitations of this study. First, the applied self-report measures have been criticized because of their cognitive and motivational drawbacks. Therefore, further studies should try to replicate the results with behavioral, physiological and performance sleepiness instruments. Secondly, sleepiness might be confounded by different level of recruiting volitional reserve capacities, i.e. different levels of effort to fight against sleepiness.

Thirdly, our results are limited by the facts that we did not consider variations in speakers´ trait (strong dialect, older age), and variations in situational context factors (e. g. noisy environments). These confounders might influence the detection rate and the false alarm error rate of the sleepiness measurement. In general, sleepy speech studies could employ a rather hybrid, i.e. data and theory driven, research paradigm rather than a pure performance optimized brute force approach. In a real-world application setting the findings of this study could be used as baseline comparison opportunity for the performance evaluation of sleepiness-detection models. In short, this work delivers critical information on the performance of large datasets in the field of acoustic sleepiness detection.

References

[1] MacLean, A.W.: Sleepiness and Driving, Sleep Medicine Reviews 7, (2003) 507-521.

[2] Melamed, S.: Excessive Daytime Sleepiness and Risk of Occupational Injuries in Non-Shift Daytime Workers", Sleep 25(3), (2002) 315-322.

[3] Wright, N., McGown, A.: Vigilance on the Civil Flight Deck: Incidence of Sleepiness and Sleep during Long-Haul Flights and Associated Changes in Physiological Parameters, Ergonomics 44, (2001) 82-106.

[4] S. Melamed, A. Oksenberg, Excessive daytime sleepiness and risk of occupational injuries in non-shift daytime workers, Sleep 25 (2002) 315–322.

[5] N. Wright, A. McGown, Vigilance on the civil flight deck: Incidence of sleepiness and sleep during long-haul flights and associated changes in physiological parameters, Ergonomics 44 (2001) 82–106.

[6] Stutts, J.C., Wilkins, J.W., Scott-Osberg, J., and Vaughn, B.V., "Driver risk factors for sleep-related crashes", Accid Anal Prev. 35, 321-331, 2003.

[7] Wierwille, W.W., "Overview of research on driver drowsiness definition and driver drowsiness detection", Proc. Enhanced Safety Vehicles (ESV) Conf., Munich, Germany, 462–468, 1994.

[8] Wright, N. and McGown, A., "Vigilance on the civil flight deck: incidence of sleepiness and sleep during long-haul flights and associated changes in physiological parameters", Ergonomics 44, 82-106, 2001.

[9] Wilhelm, H. and Wilhelm, B., "Clinical applications of pupillography", Journal Neuroophthalmol 23, 42-9, 2003.

[10] Caffier, P.P., "The spontaneous eye-blink as sleepiness indicator in patients with obstructive sleep apnoea syndrome - a pilot study". Sleep Medicine 2, 155-162, 2002.

[11] Dinges, D.F., Techniques for Ocular Measurement as an Index of Fatigue at the Basis for Alertness Management. National Highway Traffic Safety Administration, 1998.

[12] Galley, N. and Schleicher, R., Fatigue indicators from the electrooculogram -a research report. AWAKE consortium internal report, 2002.

[13] Porcu, S., 1998. "Smooth Pursuit and Saccadic Eye Movements as possible indicators of nighttime sleepiness", Physiol. Behavior. 65 (3), 437-439.

[14] Zils, E.,." Differential effects of sleep deprivation on the performance of saccadic eye movements", Sleep 28, 1109-1115, 2005.

[15] Vöhringer-Kuhnt, T., Baumgarten, T., Karrer, K., and Briest, S., "Wierwille's method of driver drowsiness evaluation revisited.", Paper at the 3rd International Conference on Traffic & Transport Psychology, 5-9, 2004.

[16] Schuller, B., Weninger, F., Wöllmer, M., Sun, Y., Rigoll, G., 2010 March. Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In: Proc. of ICASSP, Dallas, TX, USA, pp. 4562–4565.

[17] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The Interspeech 2011 speaker state challenge," in Proc. Interspeech, Florence, Italy, 2011, pp. 3201–3204.

[18] Price, C. J. (2012). A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage, 62(2), 816-847.

[19] Hudson, A. I., & Holbrook, A. (1982). Fundamental Frequency Characteristics of Young Black Adults Spontaneous Speaking and Oral Reading. Journal of Speech, Language, and Hearing Research, 25(1), 25-28.

[20] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[21] Krajewski, J., Batliner, A., & Golz, M. (2009). Acoustic sleepiness detection – Framework and validation of a speech adapted pattern recognition approach. Behavior Research Methods, 41, 795-804

[22] Krajewski, J., & Kröger, B. J. (2007). Using prosodic and spectral characteristics for sleepiness detection. In INTERSPEECH (pp. 1841-1844).

[23] Krajewski, J., Wieland, R., & Batliner, A. (2008). An acoustic framework for detecting fatigue in speech based Human-Computer-Interaction (pp. 54-61). Springer Berlin Heidelberg.

[24] Krajewski, J., Schnieder, S., Sommer, D., Batliner, A., & Schuller, B. (2012). Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. Neurocomputing, 84, 65-75.

[25] Hönig, F., Batliner, A., Bocklet, T., Stemmer, G., Nöth, E., Schnieder, S., & Krajewski, J. (2014). Are men more sleepy than women or does it only look like-Automatic analysis of sleepy speech. In ICASSP (pp. 995-999).

[26] Hönig, F., Batliner, A., Nöth, E., Schnieder, S., & Krajewski, J. (2014). Acoustic-Prosodic Characteristics of Sleepy Speech–between Performance and Interpretation. In Proc. of Speech Prosody.