Using Prosodic and Spectral Characteristics for Sleepiness Detection

Jarek Krajewski[1], *Bernd Kröger*[2]

[1] Work and Organizational Psychology, 42097 Wuppertal, Germany
[2] Clinic of Phoniatrics, Paedaudiology, and Communication Disorders, University Hospital Aachen and Aachen University
krajewski@uni-wuppertal.de, bkroeger@ukaachen.de

Abstract

This paper describes a promising sleepiness detection approach based on prosodic and spectral speech characteristics and illustrates the validity of this method by briefly discussing results from a sleep deprivation study (N=20). We conducted a within-subject sleep deprivation design (8.00 p.m to 4.00 a.m). During the night of sleep deprivation, a standardized self-report scale was used every hour just before the recordings to determine the sleepiness state. The speech material consisted of simulated driver assistance system phrases. In order to investigate sleepiness induced speech changes, a standard set of spectral and prosodic features were extracted from the sentences. After forward selection and a PCA were employed on the feature space in an attempt to prune redundant dimensions, LDA- and ANN-based classification models were trained. The best level-0 model (RA15, LDA) offers a mean accuracy rate of 80.0% for the two-class problem. Using an ensemble classification strategy (majority voting as meta-classifier) we achieved a accuracy rate of 88.2%.

Index Terms: spectral features, sleepiness detection, driver assistance system, ensemble classification

1. Introduction

Measuring sleepiness has been recognized as an important factor for the prevention of a broad range of traffic accidents [9, 10, 20, 25, 28]. 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 [26], eye blinking [2, 3, 4], eyelid movement [25], and saccade eye movement [12, 29] as well as (b) gross body movement, head movement, mannerism, and facial expression in order to characterize a driver's state of alertness [22]. In this paper we describe a spectral and prosodic approach to measure sleepiness. Our attention is focused particularly on the influence of sleepiness on driver assistance communication. The rest of this paper is organized as follows: Section 2 describes sleepiness related changes in speech. Section 3 discusses the sleep deprivation design and acoustic features used. The results of an ANN classifier are provided in Section 4, discussion and conclusions are given in Section 5.

2. Sleepiness and Speech Changes

Sleepiness related physiological changes can influence voice characteristics according to the following stages of speech production [13, 14, 16]:

(a) *cognitive speech planning*: reduced cognitive processing speed→ impaired speech planning and impaired neuromuscular motor coordination processes→ slowed articulator movement→ slackened articulation and slowed speech

(b) *respiration:* decreased muscle tension→ flat and slow respiration→ reduced subglottal pressure→ lower fundamental frequency, intensity, articulatory precision, and rate of articulation

(c) *phonation:* decreased muscle tension→ increased vocal fold elasticity and decreased vocal fold tension; decreased body temperature→ changed viscoelasticity of vocal folds

(d) *articulation/ resonance*: decreased muscle tension→ unconstricted pharynx and softening of vocal tract walls→ energy loss; postural changes→ lowered upper body and lowered head→ changed vocal tract shape; increased salivation→ energy loss; decreased body temperature→ reduced heat conduction, changed friction between vocal tract walls and air, changed laminar flows (→ energy loss)→ shift in the spectral energy distribution, broader formant bandwidth, increase in formant frequencies especially in lower formants

(e) *radiation:* decreased orofacial movement, facial expression, and lip spreading ("relaxed open mouth display") [8, 21]→ shortening of the vocal tract→ lower F1 and F2 frequencies; reduction of articulatorical effort→ smaller opening degree→ slackened articulation→ decreased first formant; oropharyngeal relaxation→ increased nasality.

However, little empirical research has been done to examine the effect of sleepiness on acoustic voice characteristics. Most studies [24, 6] have analyzed only prosody cues (i.e., intensity, speech rate, and F_0), whereas segmental cues (e.g. coded by MFCC's) have received little attention [5, 11, 14]. The aim of this study is to introduce a sleepiness detection method based on spectral and prosodic features in order to answer the questions: Can sleepiness be described quantitatively by parameters derived from segmental acoustic analysis?

3. Method

3.1. Procedure

Twenty-three students, recruited from the University of Wuppertal (Germany), took part in this study voluntarily. Initial screening excluded those having sleep disorders or sleep difficulties (PSQI). The participants were instructed to maintain their normal sleep pattern and behaviour. Due to recording and communication problems, the data of the 6 participants could not be analyzed.

We conducted a within-subject sleep deprivation design (8.00 p.m to 4.00 a.m). During the night of sleep deprivation a well proved, standardised, self-report sleepiness measure, Stanford Sleepiness Scale (SSS) [7] was used every hour just before the recordings to determine the sleepiness state. On this scale, a score of 1 point indicates “feeling active and vital, alert, wide awake” and a score of 7 points indicates “almost in reverie, sleep onset soon, losing struggle to remain awake”. The two most sleepy and the two most alert SSS measurements were included in the analysis. During the night, the subjects were confined to the laboratory and supervised throughout the whole period. Between sessions, they remained in a room, watched DVD, and talked. Non caffeinated beverages and snacks were available ad libitum.

3.2. Speech Material and Recording

The recording took place in a laboratory room with dampened acoustics using a high-quality, clip-on microphone (sampling rate: 44.1 kHz, 16 bit). The input level of the sound recording was kept constant throughout the recordings. Furthermore the subjects were given sufficient prior practice so that they were not uncomfortable with this procedure. The verbal material consisted of a German phrase, in form of a statement: “Ich suche die Friesenstraße“ [“I´m searching for the Friesen Street”]. The sentence was taken from simulated communication with a driver assistance system. The participants recorded other verbal material at the same session, but in this article we focus on the material described above [17 subjects x 4 sentences = 68 speech samples]. For training and classification purposes the records were further divided in two classes: sleepy (SS) and not sleepy (NSS) with the boundary value SSS ≥ 5. (46 samples NSS, 22 samples SS)

3.3. Feature Extraction

All acoustic measurements were taken sentence wise using the Praat speech analysis software [1]. Formant processing (F1-F5) was done with Praat, using a pre-emphasis filter with frequency response of 25 ms hamming window and 10 ms step size. For our study we estimated the following types of features:

- *prosodic features (26):* Fundamental frequency, intensity and other types of supra-segmental information such as jitter and shimmer were calculated. In particular, we computed the functionals: mean, 2.-4 quartile, standard deviation, maximum, minimum, range, positions and values of maxima and minima. Finally, we considered jitter and shimmer, short-term fluctuations in energy and fundamental frequency.

- *spectral features I (107)*: Frequencies, bandwidths, and amplitudes of the F1-F5 formants, and the frequencies and amplitudes of the first 2 harmonics. Moreover, we calculated, 4 Hammarberg indices and the average LTAS spectrum on 6 frequency bands (125-200 Hz, 200-300 Hz, 500-600 Hz, 1000-1600 Hz, 5000-8000 Hz) [16], proportion of low frequency energy under 500Hz/1000Hz, the slope of spectral energy above 1000 Hz, the Harmonic-to-Noise ratio (HNR), and spectral tilt features (“open quotient”, “glottal opening”, “skewness of glottal pulse”, and “rate of glottal closure”) [19].

- *spectral features II (36):* the usual 36 MFCC features (12 MFCC, 12 ΔMFCC, 12 ΔΔMFCC). To calculate these coefficients, we average the frame-wise computed mel-cepstral coefficients and 12 time differences over the entire signal. We expect these coefficients to account for specific properties of the sleepy speech such as increased nasalization.

3.4. Feature Selection

The purpose of feature selection is to reduce the dimensionality, which can otherwise hurt the performance of the pattern classifiers. The small amount of data also suggested that longer vectors would not be advantageous due to overlearning of data. In this study, we used stepwise linear regression method (forward selection). The threshold for adding a feature was set so that 10 (resp. 15 or 20) best features were selected (RA10, RA15, and RA20). In addition we employed Principal Component Analysis for feature selection. The first 10 (resp. 15 or 20) dimensions were extracted (PCA10, PCA15, and PCA20).

3.5. Classification

For the classification we used a *multilayer perceptron*, a special kind of artificial neural network (ANN) and a simple linear classifier (LDA). Because ANNs, specifically multi-layer perceptrons (MLPs), have proved useful for research in emotion recognition from speech, this classifier was chosen and computed with Matlab software. We used a feedforward net with backpropagation learning algorithm (one hidden layer, 5 nodes). We divided the data into training and test sets built from 34 and 34 sentences, respectively. The test set contains only speakers unseen in the training set. In the following experiments, all the classification errors were calculated by a twofold cross-validation. Using a two-fold cross-validation reduces effort and, at the same time, secures strict speaker independence. The training data was divided into two disjoint sets of equal size, and classifiers were trained twice, each time with a different set held out as a test set. The final classification errors were calculated averaging over the two test data sets. In addition to this procedure we applied an ensemble classification strategy including level-0 classifier results (see Table 1) using a majority voting as meta-classifier [27].

4. Results

We tried three different feature set sizes (10, 15, and 20) and two selection methods (linear regression based forward selection regression and PCA) to classify sleepy vs. non sleepy speech. For all configurations we trained the classifier and tested them on the test sets. The averaged accuracy rates (ratio correctly classified samples through all samples) of two different classifiers, ANN and LDA, for the two class problems are shown in Table 1.

Table 1: *Accuracy rates (in %) on the test set using different feature set size (10, 15, and 20 feature), different feature set selection method (stepwise linear regression analysis or principal component analysis) and different classifier (linear discriminant analysis or artifical neural net).*

		RA	**PCA**
LDA	10	70.6	75.7
	15	80.0	76.4
	20	75.0	67.7
ANN	10	76.4	58.8
	15	79.4	76.4
	20	70.6	53.0

Within the best 20 features the following numbers of features were included: 2 prosodic, 7 spectral I, and 11 spectral II feature (MFCCs). The best results were achieved with the RA15 feature set with LDA classifier (80.0%) and the RA15 feature set with ANN classifier (79.4%). Ensemble generalization was proposed for a classification task by combining multiple models. To maximize the classification accuracy one should use ensemble classification rather than any single classifier by itself. In contrast to stacking strategy we use the output of different feature subsets rather than different classifier as level-1 features. The majority voting was selected as meta classification rule. The speech sample was classified as sleepy if the propotion of the level-0 classifier output predicted as sleepy was larger than 33.3%.

Table 2: *Accuracy rates (in %) using different ensemble classification strategies for LDA: aggregation of different feature set sizes (10, 15, 20) and selection methods (RA, PCA).*

	RA	**PCA**
Voting (10,15,20)	85.3	76.4
Voting RA+PCA (10,15,20)	88.2	
Voting RA+PCA 10	85.3	
Voting RA+PCA 15	85.3	
Voting RA+PCA 20	88.2	

Table 2 shows the performance comparison between different level-0 feature sets. The combination of all 6 level-0 classifier feature sets (RA10, RA15, RA20, FA10, FA15, FA20) perform best (recognition rate 88.2%). A more detailed look on the best classification result is presented in Table 3, where the confusion matrix is depicted. The proportion of correct classified sleepy speech is 72.7% (detection rate). On the other hand 4.3% false alarm errors can be found. The kappa coefficient which thought to be the chance corrected prediction accuracy is .72 [kappa=0.88- 0.57)/ (1-0.57)].

Table 3: *Confusion matrix on the test set using the voting meta classifier with 6 LDA-based* level-0 feature sets. [SS = sleepy and NSS = not sleepy].

	Hypothesis	
Reference	**NSS**	**SS**
NSS	44 (95.7%)	2 (4.3%)
SS	6 (27.3%)	16 (72.7%)

5. Discussion

A crucial aim of this study was to explore whether voice features are associated with sleepiness. The main findings of the present study may be summarised as follows. First, prosodic and spectral features extracted from driver assistance system communication contain a different amount of information about sleepiness states. Within the best 20 features 2 prosodic, and 18 spectral features were selected by the forward selection. Secondly, in our experiments on a two-class classification problem (sleepy vs. non sleepy speech), we achieved a accuracy rate of 80.0% on unseen data. The best recognition performance is attained for a 15 feature set using a LDA classifier. Thirdly the ensemble classification strategy (majority votings as meta classifier) using the output of the level-0 classifier offered a recognition rate of 88.2%. The accuracy rate increased in comparison to the best level-0 classifier by 8.2%. From the confusion matrix it is evident that the meta classifier perfoms with a detection rate of 72.7% and a false alarm rate of 4.3%.

Due to the hypothezised sleepiness related physiological changes in cognitive speech planning, respiration, phonation, articulation, and radiation, the results for the reported classification performance were largely as could be expected. This is consistent with previous sleepiness related findings, that suggest an association of prosodic [24, 6] and spectral characteristics [5, 11, 14] with sleepiness.

Limitations. 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 behavioural, physiological and performance sleepiness instruments. Secondly, sleepiness might be confounded by annoyance states due to the multiple repetition of speak task. Thus the results obtained in the current study with a within subject design should be replicated with a between subject design. Thirdly, our results are limited by the facts that we did not consider real life speaking conditions including variation in speakers´ states (having a headcold, drinking milk, being nervous, aggressive or in a depressive mood), variations in speakers´ trait (strong dialect, older age), and variations in situational context factors (high driving related work-load situation, noisy viroments). These confounders might influence the detection rate and the false alarm error rate of the sleepiness measurement. Furthermore the analysis assumes a closely placed microphone and noise-free recordings of short sentences. However, it is not realistic to expect such a clean audio input, especially not in unconstrained traffic environments in which automatic sleepiness detection systems are most likely to be deployed.

Future work. Granted, the present results are preliminary and need to be replicated using natural speech enviroment. Nevertheless, it would seem advisable that future studies address the following topics:

- segmentation: finding sleepiness sensitive phonetical units (phones or VCV in different word and phrasal unit position)
- feature extraction: computing of rhythm and duration related features; time-domain based features from nonlinear time series analysis (lyapunov exponents, correlation dimension, automutual information, time resolved density, fractal dimensions, multiscale entropies, and recurrence quantification analysis [23]); using automatic feature generation
- pattern classification: dividing between male and female classification models; utilizing SVMs, maximum-likelihood bayes classifiers, kNNs, fuzzy membership indexing, ANNs, HMMs, gaussian mixture density models; using sophisticated ensemble classification methods (boosting, stacking) [17].

6. References

[1] P. Boersma, P., “PRAAT, a system for doing phonetics by computer”, Glot International, vol. 5, no. 9/10, pp. 341–345, 2001.

[2] 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.

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

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

[5] Greely, H., “Fatigue Prediction Using Voice Analysis”, Behavioral Research Methods, 2007.

[6] Harrison, Y. and Horne, J.A., “Sleep deprivation affects speech”, Sleep, 20, 871-877, 1997.

[7] Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., and Dement, W.C., Stanford Sleepiness Scale,1973.

[8] Kienast, M. and Sendlmeier, W.F., “Acoustical analysis of spectral and temporal changes in emotional speech”, Speech Emotion, 92-97, 2000.

[9] MacLean, A.W., “Sleepiness and driving”, Sleep Medicine Reviews 7, 507-521, 2003.

[10] Melamed, S., “Excessive daytime sleepiness and risk of occupational injuries in non-shift daytime workers”, Sleep 25(3), 315-322, 2002.

[11] Nwe, T.L., Li, H., and Dong, M., “Analysis and Detection of Speech under Sleep Deprivation”, Interspeech 17-21, 2006.

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

[13] Rabiner, L. and Schafer, R.W., Digital Processing of Speech Signals, Prentice-Hall, Upper Saddle River, New Jersey, USA, 1978.

[14] Raghavan, S., Applications of large vocabulary continuous speech recognition to fatigue detection. Mississippi State University, 2006.

[15] Scherer, K.R., “Vocal affect expression: A review and a model for future research”, Psychological Bulletin 99, 143-165, 1986.

[16] Scherer, K.R., Johnston, T., and Klasmeyer, G., Vocal Expression of Emotion. In R.J. Davidson, K.R. Scherer, H.H. Goldsmith (Eds). Handbook of Affective Sciences, 433-456, 2003.

[17] Schuller, B., Automatische Emotionserkennung aus sprachlicher und manueller Interaktion, Technische Universität München, 2006.

[18] Stevens, K., Acoustic Phonetics. MIT Press. Cambridge, England, 1999.

[19] Stevens, K. and Hanson, H., “Classification of glottal vibartion from acoustic measurements”, Vocal Fold Physiology, 147-170, 1994.

[20] 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.

[21] Tartter, V.C., “Happy talk: Perceptual and acoustic effects of smiling on speech”. Perception and Psychophysics 27 (1), 24-27, 1980.

[22] 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.

[23] Webber, C. L., Zbilut, J. P., “Dynamical assessment of physiological systems and states using recurrence plot strategies”, Journal of Applied Physiology 76, 1994.

[24] Whitmore, J. and Fisher, S., “Speech during sustained operations”, Speech Communication 20, 55-70, 1996.

[25] 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.

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

[27] Wolpert, D.H., “Stacked Generalization”, Neural Networks 5, 241-259, 1992.

[28] 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.

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