Acoustic-Prosodic Characteristics of Sleepy Speech – Between Performance and Interpretation*

Florian Hönig[1], *Anton Batliner*[1,2], *Elmar Nöth*[1,3], *Sebastian Schnieder*[4], *Jarek Krajewski*[4]

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]Institute for Human-Machine Communication, Technische Univ. München, Munich, Germany
[3]Electrical & Computer Engineering Dept., King Abdulaziz University, Jeddah, Saudi Arabia
[4]Experimental Industrial Psychology, University of Wuppertal, Germany
{hoenig,batliner}@cs.fau.de

Abstract

When we address speaker states like sleepiness, two partly competing interests can be observed: both within applications and engineering approaches, we aim at utmost performance in terms of classification or regression accuracy – which normally means using a very large feature vector and a brute force approach. The other interest is interpretation: we want to know what tells apart atypical (here: sleepy) speech from typical (here: non-sleepy) speech, i.e., their respective feature characteristics. Both interests cannot be served at the same time. In this paper, we preselect a small number of easily interpretable acoustic-prosodic features modelling spectrum and prosody, based on the literature and on the general idea of sleepiness being characterised by relaxation. Performance obtained with these single features and this small feature vector is compared with the performance obtained with a very large feature vector; moreover, we discuss to which extent the features chosen model relaxation as sleepiness characteristic.

Index Terms: paralinguistics, sleepiness, prosody, brute forcing, interpretation

1. Introduction

Sleepiness is definitely an interesting research topic, both for practical reasons – the detection of sleepiness is highly relevant in scenarios where sleepiness can cause accidents (driving, flying, operating of machines), and for general reasons – it is ubiquitous, we face it several times a day. As a multi-modal phenomenon, it can be perceived/measured within all modalities, be this speech, facial gestures, eye movements, gait, body posture, or biosignals. Each of these modalities has its pros and cons, as far as processing is concerned: for video processing, light conditions should be favourable; biosensors are intrusive; audio recordings are non-intrusive and possible even under less favourable noise conditions. In this paper, we will concentrate on audio. Moreover, we concentrate on one specific research problem which might not be formulated that often explicitly but sort of gets into our way very often: do we want to get better, or do we want to get any wiser? For getting better, i.e., for obtaining the highest accuracy in classification or regression, we normally employ a very large feature vector (with or without subsequent feature selection): the baseline result for the Interspeech 2011 sleepiness challenge [1] was obtained using the openSMILE tool and 4368 features. Finding features that are most relevant for performance and at the same time easily interpretable is not an easy task [2]. For getting any wiser, so far, we are confined to the rather 'traditional' way of doing research: we employ a small set of promising features and, if possible, formulate a working (alternate) hypothesis on what we expect to find. These promising features are at the same time easily interpretable such as *F0 mean*, in contrast to complex and at the same time opaque features such as the *75% quantile of the 10th MFCC coefficient on consonantal frames* which will turn out as the 2nd most important feature obtained in our data-driven feature selection, cf. Section 4. However, we will definitely not get the highest possible classification/regression performance when using the 'traditional' approach. Here, we try to combine these two different approaches. After presenting the database in Section 2, we sketch in Section 3 the feature sets employed. The experiments reported in Section 4 are discussed in Section 5.

2. Data and Annotation

We employ the Sleepy Language Corpus (SLC) from the Interspeech 2011 Speaker State Challenge [1, 3]. Ninety-nine German speakers took part in six partial sleep deprivation studies (mean age 24.9 years, standard deviation 4.2 and a range of 20–52 years; recordings in a realistic car environment or in lecture-rooms; microphone-to-mouth distance 0.3 m, sampling rate 16 kHz, quantisation 16 bit). We disregard the isolated vowels and use the remaining five subsets (7745 speech files ("turns", units of analysis in our regression approach), about 20 hours of speech): read speech: the story of "Die Sonne und der Nordwind" ('the North Wind and the Sun'); commands/requests: simulated driver assistance system commands/requests, e. g. "Ich suche die Friesenstraße" ('I am looking for the Friesen street'); simulated pilot-air traffic controller communication statements (non-native English); descriptions of pictures; a PowerPoint guided, but non-scripted 20 minutes presentation in front of 50 listeners. A well established, standardised subjective sleepiness questionnaire, the Karolinska Sleepiness Scale (KSS, [4]), 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. Scores range from 1 to 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, strug-

* The authors have received funding from the German Research Council (DFG) under grant agreements KR 3698/4-1 and NO 444/6-1. The responsibility lies with the authors.

10.21437/SpeechProsody.2014-161

gling against sleep (9), extremely sleepy, cannot stay awake (10). The labels were given not to single turns but to 'recording units' consisting of up to 20 turns (9.4 on average) such as stories or sequences of commands. This constitutes an optimal and smooth reference; accordingly, mean pairwise Pearson correlation between self-assessment and observers is very high: 0.89 (0.88 between two observers). The scores from self-assessment and observers are averaged to form the reference sleepiness values (mean/standard dev.: 6.1 ± 2.3 for females, 5.9 ± 2.5 for males). A more detailed description of the data is given in [5, 6]. For the 2011 Challenge, a subdivision of the data into three speaker-disjunct sets for training, development and test was defined. Here, we always report the results on the test set (*TEST*, 19 females, 14 males, 2466 turns, 6.6 hours), estimating parameters on the union of the original training and development set (henceforth *TRAIN*, 37 females, 29 males, 5279 turns, 13.2 hours). Gender is a bit imbalanced: 73% and 64% of the utterances of TRAIN and TEST, respectively, are from female speakers.

3. Features

We employ 3705 acoustic-prosodic features described in [7]; here, we only can give the general idea. For segmenting pauses, vowels, consonants, and speaker noise, we use the phoneme recognizer of the Brno Univ. of Technology [8]. Then, pseudo-syllables are derived in four different ways, taking: (1) the nucleus (i. e. consecutive vowels), (2) nucleus + coda (consecutive vowels plus trailing consecutive consonants), (3) onset + nucleus (leading consonants plus consecutive vowels), and (4) onset + nucleus + coda (leading consonants plus consecutive vowels plus trailing consonants – these syllables overlap). We compute four low-level descriptors on a frame-by-frame basis: F0, formants, formant bandwidths, and Mel frequency cepstral coefficients (MFCC) as a more fine-grained and robust, yet less explicit representation of articulators. For each syllable, we compute micro-structural prosodic descriptors such as loudness [9]. F0 is suitably interpolated, normalized per utterance, and perceptually transformed. Normalized versions of energy and duration remove phoneme-intrinsic influences. To obtain a fixed number of features per utterance, we compute twelve functionals that characterise the statistical and temporal properties of these local descriptors: mean, standard deviation, minimum, maximum, median, quantiles 5%, 25%, 75%, 95%, average absolute local change (similar to Grabe's raw pairwise variability index rPVI [10]), root average squared local change, and slope of the regression line. Depending on the type of descriptor, these functionals are computed across syllables and vocalic/consonantal frames. Additionally, we compute features developed for describing speech rhythm [10, 11, 12]. From this brute-force set, we now manually select and combine the following features suitable to capture the acoustic correlates of sleepiness that can be expected according to the pertinent literature (see e. g. [3] for an overview). The expected sign given in parentheses indicates falling/lower ('$-$') or rising/higher ('$+$') values for sleepy speech; an appended question mark indicates ambiguous tendencies.

spectral features: formants

(1) g-mean(F1–4)_V_mean ($-$): the ***geometric mean of formants F1–F4*** per frame, averaged across vocalic frames. The circadian rhythm includes body core temperature variation [13]; we can assume some decrease with sleepiness. With that, the temperature of the exhaled air drops, too. Therefore, formant frequencies should be shifted slightly downwards [14, 15].

(2) mean(FBW1–4)_V_mean ($+$): the ***arithmetic mean of the formant bandwidths FBW1–FBW4 per frame***, averaged across vocalic frames. Reduced body temperature and muscular relaxation might lead to vocal tract softening and stronger dampening of the signal due to yielding walls [16]. We expect glottal loss and cavity-wall loss for the lower formants, and radiation, viscous and heat-conduction loss for the higher formants [17]. Consistent with that is the increased time of high values for Formant 1 bandwidth in [6].

(3) F1_V_std * F2_V_std ($-$): the ***product of the standard deviations of F1 and F2*** across vocalic frames. The reduced cognitive processing speed going along with sleepiness might lead to impaired neuromuscular motor coordination processes, slowing down the transduction of neuromuscular commands into articulator movement and affecting the feedback of articulator positions [18, 19], possibly leading to aversion of spending compensatory effort [20]. Thus, sleepy speech could exhibit slurred, less crisp pronunciation, mispronunciations, abrupt articulatory changes, speech errors, or hesitations. A less crisp pronunciation might result in vowel centralization and a reduced area covered by the first two formant frequencies, which account for most of the discriminability across vowels.

(4) F1_V_mean ($-$?): the ***average of F1 across vocalic frames***. Sleepy speech is also expected to exhibit changes in speech quality such as tensed, nasal, or breathy speech due to, e. g. impaired coordination of velum closure [21]. The effects of increased nasality are complex: the first formant (F1) gets weaker, and its position moves higher, because nasals are usually pronounced more open. Yet, F1 is likely to be masked by the appearance of the lower and louder first nasal formant, resulting in an opposite tendency. Thus, nasality seems to be difficult to quantify with a simple acoustic parameter. The decrease in F1 for sleepy speech reported in [6] thus cannot be readily assessed: most likely, the decrease is due to the first nasal formant showing up more, and possibly also to reduced body temperature as shown above, these two effects outweighing the opposite tendency due to the expected more open pronunciation.

spectral features: MFCC

(5) MFCC2_V_mean ($+$?): the ***average of the second MFCC coefficient*** across vocalic frames as an estimate of the negative spectral tilt; we expect the spectral tilt to fall with sleepiness, and thus a rise of this feature. Increased breathiness in sleepy speech – see feature (4) – should lead to a negative spectral tilt for high frequencies [22], which seems to be confirmed by [23] where a decrease of the slope of the long term average spectrum is reported. An opposite effect could be caused, however, by a stronger high-pass effected by a more closed mouth position (centralisation) compatible with reduced muscular tension.

(6) MFCC1_V_mean / MFCC1_C_mean ($-$): the ***ratio of the first MFCC coefficient averaged across vocalic frames to its average over consonantal frames.*** The abovementioned losses in resonance (muscular relaxation and reduced body temperature) could lead to a reduced energy in vocalic segments compared to consonantal segments; the first MFCC coefficient is a measure of energy.

(7-10) MFCC2_V_std, MFCC3_V_std, MFCC2_C_std, MFCC3_C_std ($-$): the ***standard deviations of the second and third MFCC computed separately across vocalic and consonantal segments.*** These features describe coarsely the spectrum of the vowels (F1, F2) and of the consonants and can capture less diligent pronunciation (centralisation) [7].

prosody: F0

(11) F0_V_mean ($-$?): the ***average of pitch estimates across vocalic frames.*** The muscular relaxation going along with sleepiness might lead to a reduced fundamental frequency (F0) as reported in [6, 24, 25], although [26] report the opposite. Since increased breathiness should also go along with reduced F0, there is one more reason to assume a decrease.

(12) F0_V_std / F0_V_mean ($-?$): the ***standard deviation of F0, normalized to the mean F0***, across vocalic frames. For sleepy speech, we anticipate monotonic and flattened intonation [3]: [27, 25] report a decreased standard deviation of F0 although an opposite result has been published in [28]. The standard deviation of pitch is correlated with the absolute pitch level of a speaker; this effect is removed by the normalization.

(13) syl-F0-mean_std ($-?$): the ***standard deviation of the syllables' average F0***; here and in the following, we use the 'nucleus + coda' pseudo-syllables. Now we apply our microstructural prosodic features, where F0 undergoes a different normalization, and perceptual scaling.

(14) syl-F0-max_mean ($-?$): the ***syllables' F0 maxima*** averaged across syllables. Flattened intonation should lead to less pronounced F0 maxima.

(15) syl-F0-min_mean ($+?$): the ***syllables' F0 minima*** averaged across syllables which should rise.

(16) syl-F0-slope_mean ($-$): the ***F0 slope within syllables***, averaged across syllables, expected to fall with sleepiness because of flattened intonation.

prosody: energy

(17) syl-energy-mean-norm_std ($+?$): the ***standard deviation of the syllables' normalized mean energy.*** According to [6], the average absolute deviation of intensity increases with sleepiness. This could be explained as a less diligent or controlled pronunciation, although a flattened intonation might also have the opposite effect.

(18) syl-rel-energy_mean ($-$): a ***medium-term estimate of the relative energy*** (computed for energy normalization purposes [9] from up to 15 neighbouring syllables, taking into account phoneme-intrinsic properties), averaged across syllables. Muscular relaxation might also lead to reduced loudness.

(19) syl-energy-slope_mean ($-$): the ***average energy slope within syllables*** which we expect to fall with sleepiness, due to flattened intonation.

prosody: duration

(20) syl-rel-duration_mean ($+$): ***medium-term estimates of the syllables' relative durations***, averaged across syllables. Slowed cognitive processing reduces speech planning, which might lead to a reduced speech rate [29, 27, 28] and thus to increased durations.

(21-22) syl-pauses_mean, syl-filled-pauses_mean ($+$): the ***average duration of silent and of filled pauses between syllables.*** Along with segment durations, pause length is expected to increase with sleepiness, too [28].

prosody: rhythm

(23) %V ($-$): ***Ramus' %V, the percentage of vocalic intervals*** [11] is expected to fall because the relative frequency of voicing decreases with sleepiness [30].

(24-25) nPVI_V, nPVI_C ($-?$): ***Grabe's normalized pairwise variability index nPVI*** [10], a rate-of-speech-normalized measure of local durational variability, computed separately for vocalic and consonantal segments, is expected to fall with monotonicity (although more disfluencies could also lead to a rise).

(26-27) varco_V, varco_C ($-?$): ***Dellwo's variation coefficients*** [12] are a measure of global durational variability (rate-of-speech-normalized standard deviations of the duration of vocalic and consonantal segments). Again, we expect a decrease, although disfluencies could have the opposite effect.

4. Experiments and Results

4.1. Analysis of single Features

We compute Pearson's correlation coefficient r between the reference sleepiness values and the individual features of each utterance only for TEST; this guarantees strict comparability with the regression results of Section 4.2. Spearman's ρ did not differ much, so we skip it. The results are given in Table 1. These individual correlations are mostly weak; for the weakest correlations, contra-intuitive effects can arise. For instance, feature (3) is negatively correlated to sleepiness for female and male speakers separately, but positively for all speakers together – this can be due to slightly different distributions of feature range and sleepiness score for female vs. male speakers. For males, correlations are mostly stronger (average $|r|$: 0.14 vs. 0.09). Using the same database, we showed in [7] that this can mainly be attributed to females showing their sleepiness less than males do. If we disregard very weak correlations – arbitrarily defined as $|r| < 0.1$ – which might well be caused by noise, given the limited number of speakers – then only 9 out of 81 cases with 'unexpected' sign remain (typeset in italics in Table 1); thus, our predictions of Section 3 are generally corroborated. The correlations of feature (5) are negative, contradicting our expectation; a more relaxed and thus closed mouth position might outweigh the effects of breathiness. Feature (19) displays highly contradicting signs as well. Our conjecture was that flattened intonation would result in negative slopes, for both F0 and energy within syllables. This did turn out right for F0, see feature (16), but not for energy. Feature (2), an estimate of the bandwidth of formants, unexpectedly falls with increased sleepiness for male speakers. One conjecture would be interactions between changes in F0 and formant extraction; but then, these interactions should be stronger for females due to the higher distance between harmonics. Another explanation could be a stronger volitional effort to fight against sleepiness in women, which might lead to muscular tension and vocal tract hardening.

Feature (6), the ratio of the energy of voiced and unvoiced segments, unexpectedly rises with sleepiness for male speakers. An explanation could be a less diligent control of the air stream, possibly resulting in louder vowels for males who tend to show the effects of sleepiness to a higher extent than females [7]. Feature (18) unexpectedly rises for female speakers: the tendency of females to show sleepiness to a lesser extent might lead to overcompensation, resulting in louder speech with clearly articulated consonants (because of the negative sign for (6)).

4.2. Regression Experiments

For robust estimation, ridge regression [31] is used. Parameters are estimated on TRAIN, results are computed on TEST; details are given in [7]. Again, Spearman's ρ is similar to Pearson's r between predicted and reference sleepiness values, so we use only the latter. The results are given in Table 2. If all 3705 features are used, the best result is 0.41 for all speakers. Also here we see higher correlations for male speakers (0.50 vs. 0.34), consistent with the single correlations above.

For the 27 manually selected features, correlations are lower, e. g. 0.33 vs. 0.41 for all speakers, but much better than a pure random selection of 27 features, which results in a correlation of 0.15 for all speakers on average (not displayed in Table 2). When looking at spectral and prosodic features separately, there is another interesting gender effect: for men, spectral features seem to be more suited than prosodic features (0.43 vs. 0.21). It is the other way around for female speakers: here, prosodic features yield better results than spectral features (0.33

Table 1: *Manually selected features and their Pearson correlation r to sleepiness: for all speakers, females (f), and males (m). The absolute value of the correlations is illustrated by the grey level of each cell's background. Grossly unexpected correlations (different sign and $|r| \geq 0.1$) are set in italics.*

Feature	exp. sign	all	f	m
(1) g-mean(F1–4)_V_mean	−	+0.06	−0.09	−0.22
(2) mean(FBW1–4)_V_mean	+	+0.12	+0.11	*−0.21*
(3) F1_V_std * F2_V_std	−	+0.03	−0.04	−0.16
(4) F1_V_mean	− ?	−0.08	−0.18	−0.24
(5) MFCC2_V_mean	+ ?	*−0.19*	−0.07	*−0.41*
(6) MFCC1_V_mean/MFCC1_C.	−	−0.11	−0.21	*+0.16*
(7) MFCC2_V_std	−	−0.03	+0.01	−0.23
(8) MFCC3_V_std	−	−0.18	−0.17	−0.35
(9) MFCC2_C_std	−	+0.02	+0.01	−0.01
(10) MFCC3_C_std	−	−0.14	−0.10	−0.28
(11) F0_V_mean	− ?	+0.05	−0.26	−0.12
(12) F0_V_std / F0_V_mean	− ?	−0.02	+0.03	−0.18
(13) syl-F0-mean_std	− ?	−0.03	−0.06	+0.04
(14) syl-F0-max_mean	− ?	−0.03	−0.04	+0.04
(15) syl-F0-min_mean	+ ?	−0.02	+0.00	−0.09
(16) syl-F0-slope_mean	−	−0.08	−0.06	−0.06
(17) syl-energy-mean-norm_std	+ ?	+0.02	+0.01	+0.08
(18) syl-rel-energy_mean	−	*+0.15*	*+0.14*	−0.11
(19) syl-energy-slope_mean	−	*+0.17*	*+0.22*	*+0.10*
(20) syl-rel-duration_mean	+	+0.23	+0.22	+0.23
(21) syl-pauses_mean	+	+0.01	−0.03	+0.10
(22) syl-filled-pauses_mean	+	+0.21	+0.20	+0.12
(23) %V	−	−0.05	−0.02	−0.01
(24) nPVI_V	− ?	−0.07	−0.06	−0.10
(25) nPVI_C	− ?	+0.03	+0.04	+0.03
(26) varco_V	− ?	−0.09	−0.06	−0.17
(27) varco_C	− ?	+0.01	+0.01	−0.01

vs. 0.20). As for the non-ambiguous features ('−' or '+' without '?' in Table 1), results for all speakers suffer only a little by this restriction (0.30 vs. 0.33). However, now there is hardly a difference between the performance on female and male speakers (0.29 and 0.30). This is quite different when looking at the features we just removed: Training only with the 12 ambiguous features ('−?' or '+?' in Table 1), the performance difference between male and female speakers is more pronounced than ever (0.42 vs. 0.19). A possible explanation for this could be the following: the non-ambiguous features generally model sleepiness changes based on 'physiological primitives' that cannot be controlled very well by the speaker. The ambiguous features, where we identified possible antagonistic influences, however, represent parameters where the speakers do have some choice.

For a data-driven feature selection, we use a so-called wrapper approach, together with a greedy forward search: each time that feature is added which yields the best performance when training and testing the regression system with TRAIN. Here, we discuss the comparable numbers of selected features, namely 27 and 15, respectively. Intriguingly, these yield similar performance compared to the manual feature selection: for 27 features and all speakers, 0.33; for females and males separately, correlations decrease slightly (0.30 vs. 31, and 0.35 vs. 0.40, respectively). The first 15 automatically selected features are slightly better than the 15 manual non-ambiguous features (0.32/0.30/0.35 vs. 0.30/0.29/0.30). One would normally expect data-driven selection to outperform manual selection; however, it has to cope with weak sleepiness effects, facing noisy data from a limited number of speakers, and thus the unavoidable train-test mismatch. In fact, the 9th and the 21th selected feature are our manual features (6) and (10) – a very nice outcome because the probability for this to happen by chance is very low. Generally, the data-driven and manual features are not very similar, at least when compared individually: mean pairwise Pearson correlation between data-driven and manual features is 0.04, mean absolute 0.14. Minimal correlation is -0.53; maximal correlation (apart from the two identical features) is 0.91: between the 95%-quantile of pitch estimates across vocalic frames and our manual feature (11).

Table 2: *Performance when predicting sleepiness from different features. Both male and female speakers were used in training; Pearson correlation on test is reported for all, female (f), and male speakers (m). Higher absolute correlation = darker.*

Features	all	f	m
all (3705)	0.41	0.34	0.50
manually selected (27)	0.33	0.31	0.40
– spectral (10)	0.29	0.20	0.43
– prosodic (17)	0.22	0.33	0.21
manually selected – non-ambig. (15)	0.30	0.29	0.30
– spectral (8)	0.21	0.17	0.26
– prosodic (7)	0.30	0.32	0.21
manually selected – ambiguous (12)	0.16	0.19	0.42
– spectral (2)	0.19	0.16	0.40
– prosodic (10)	0.03	0.15	0.16
data-driven selection of 27	0.33	0.30	0.35
data-driven selection of 15	0.32	0.30	0.35

5. Discussion and Concluding Remarks

Expectably, brute forcing with many features beats knowledge-based selection of features (overall performance not being too high, obviously because sleepiness can only be partly modelled by speech alone, and its indication is partly speaker-dependent/idiosyncratic). However, our knowledge-based vector is on par – and in a few cases, overlapping – with the same number of automatically selected most important features, corroborating our general hypothesis of sleepiness being a relaxation phenomenon. However, females and males display interesting and partly antagonistic tendencies: male sleepiness is mainly reflected by spectral changes towards less canonical pronunciation (centralisation, cf. the MFCC features in Table 1) whereas female sleepiness primarily implies prosodic changes such as lowered pitch (feature 11). All this is in line with our explanation in [7], cf. [32, p. 130] and [33], that women tend towards more canonical speech. Generally, the non-ambiguous 15 features seem to be more 'stable' and more uniformly used by both males and females; in contrast, the 12 ambiguous features (esp. the spectral ones) obviously offer more degrees of freedom, e.g. for females, to 'hide', and for males, to 'express' their sleepiness. Of course, these explanations are tentative and have to be corroborated with future studies and additional data.

6. References

[1] 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.

[2] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, “Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech,” *Computer Speech and Language, Special Issue on Affective Speech in real-life interactions*, vol. 25, no. 1, pp. 4–28, 2011.

[3] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, “Medium-term speaker states – A review on intoxication, sleepiness and the first challenge,” *Computer Speech and Language*, vol. 27, pp. 1–30, 2013.

[4] A. Shahid and K. Wilkinson, “Karolinska sleepiness scale (KSS),” in *STOP, THAT and One Hundred Other Sleep Scales*, A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, Eds. Springer, 2012, pp. 209–210.

[5] J. Krajewski and B. Kroeger, “Using prosodic and spectral characteristics for sleepiness detection,” in *Proc. Interspeech*, Antwerp, 2007, pp. 1841–1844.

[6] J. Krajewski, A. Batliner, and M. Golz, “Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach,” *Behavior Research Methods*, vol. 41, pp. 795–804, 2009.

[7] F. Hönig, A. Batliner, T. Bocklet, G. Stemmer, E. Nöth, S. Schnieder, and J. Krajewski, “Are men more sleepy than women or does it only look like – automatic analysis of sleepy speech,” in *ICASSP 2014, International Conference on Acoustics, Speech, and Signal Processing, May 4-9, 2014, Florence, Italy, Proceedings*, 2014, to appear.

[8] P. Schwarz, P. Matejka, and J. Cernocky, “Hierarchical structures of neural networks for phoneme recognition,” in *Proc. of ICASSP*, Toulouse, 2006, pp. 325–328.

[9] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, “The Prosody Module,” in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Berlin: Springer, 2000, pp. 106–121.

[10] E. Grabe and E. L. Low, “Durational variability in speech and the rhythm class hypothesis,” in *Laboratory Phonology VII*, C. Gussenhoven and N. Warner, Eds. Berlin: de Gruyter, 2002, pp. 515–546.

[11] F. Ramus, “Acoustic correlates of linguistic rhythm: Perspectives,” in *Proc. Speech Prosody*, Aix-en-Provence, 2002, pp. 115–120.

[12] V. Dellwo, “Influences of speech rate on the acoustic correlates of speech rhythm. An experimental phonetic study based on acoustic and perceptual evidence,” Ph.D. dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, 2010.

[13] J. Zulley, R. Wever, and J. Aschoff, “The dependence of onset and duration of sleep on the circadian rhythm of rectal temperature,” *Pflügers Archiv*, vol. 391, no. 4, pp. 314–318, 1981.

[14] E. G. Richardson, *Technical Aspects of Sound: Sonic range and airborne sound*. Elsevier, 1953.

[15] F. Reif, *Fundamentals of Statistical and Thermal Physics*. Waveland, 2008.

[16] T. Ananthapadmanabha, “Aerodynamic and acoustic theory of voice production,” in *Forensic speaker recognition, law enforcement and counter-terrorism*, A. Neustein and H. A. Patil, Eds. New York: Springer, 2011, pp. 309–363.

[17] B. Story, “An overview of the physiology, physics and modeling of the sound source for vowels,” *Acoustical Science and Technology*, vol. 23, pp. 195–206, 2002.

[18] D. Bratzke, B. Rolke, R. Ulrich, and M. Peters, “Central slowing during the night,” *Psychological Science*, vol. 18, pp. 456–461, 2007.

[19] D. Dinges and N. Kribbs, “Performing while sleepy: Effects of experimentally-induced sleepiness,” in *Sleep, Sleepiness and Performance*, T. Monk, Ed. Chicester, England: Wiley, 1991, pp. 97–128.

[20] P. Lieberman, B. G. Kanki, and A. Protopapas, “Speech production and cognitive decrements on Mount Everest,” *Aviation, Space, and Environmental Medicine*, vol. 66, pp. 857–864, 1995.

[21] B. E. Kostyk and A. Putnam Rochet, “Laryngeal airway resistance in teachers with vocal fatigue: A preliminary study,” *Journal of voice*, vol. 12, no. 3, pp. 287–299, 1998.

[22] D. H. Klatt and L. C. Klatt, “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” *the Journal of the Acoustical Society of America*, vol. 87, p. 820, 1990.

[23] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, and B. Schuller, “Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech,” *Neurocomputing, Special Issue “From neuron to behavior: evidence from behavioral measurements”*, vol. 84, pp. 65–75, 2012.

[24] B. Johannes, V. P. Salnitski, H.-C. Gunga, and K. Kirsch, “Voice stress monitoring in space – possibilities and limits,” *Aviation, Space, and Environmental Medicine*, vol. 71, pp. A58–65, 2000.

[25] T. L. Nwe, H. Li, and M. Dong, “Analysis and detection of speech under sleep deprivation,” in *Proc. of Interspeech*, 2006, pp. 17–21.

[26] R. Ruiz, P. Plantin De Hugues, and C. Legros, “Advanced voice analysis of pilots to detect fatigue and sleep inertia,” *Acta Acustica united with Acustica*, vol. 96, pp. 567–579, 2010.

[27] G. O. Morris, H. L. Williams, and A. Lubin, “Misperception and disorientation during sleep deprivation,” *Archive of General Psychiatry*, vol. 2, pp. 247–252, 1960.

[28] A. P. Vogel, J. Fletcher, and P. Maruff, “Acoustic analysis of the effects of sustained wakefulness on speech,” *Journal of the Acoustical Society of America*, vol. 128, pp. 3747–3756, 2010.

[29] W. J. M. Levelt, A. Roelofs, and A. S. Meyer, “A theory of lexical access in speech production,” *Journal of Behavioral and Brain Sciences*, vol. 22, pp. 1–75, 1999.

[30] L. Dhupati, S. Kar, A. Rajaguru, and A. Routray, “A novel drowsiness detection scheme based on speech analysis with validation using simultaneous EEG recordings,” in *Proc. IEEE Conference on Automation Science and Engineering (CASE)*, Toronto, ON, 2010, pp. 917–921.

[31] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, pp. 55–67, 1970.

[32] J. Kreiman and D. Sidtis, *Foundations of Voice Studies - An Interdisciplinary Approach to Voice Production and Perception*. Wiley, 2011.

[33] P. Trudgill, “Sex, Covert Prestige and Linguistic Change in the Urban British English of Norwich,” *Language in Society*, vol. 1, pp. 175–195, 1972.