ARE MEN MORE SLEEPY THAN WOMEN OR DOES IT ONLY LOOK LIKE – AUTOMATIC ANALYSIS OF SLEEPY SPEECH*

Florian Hönig[1], *Anton Batliner*[1,2], *Tobias Bocklet*[1], *Georg Stemmer*[1], *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}@informatik.uni-erlangen.de

ABSTRACT

The degree of sleepiness in the Sleepy Language Corpus from the Interspeech 2011 Speaker State Challenge is predicted with regression and a very large feature vector. Most notable is the great gender difference which can mainly be attributed to females showing their sleepiness less than males do.

Index Terms— Sleepiness, paralinguistics, regression, brute forcing, gender differences

1. INTRODUCTION

Sleepiness can be a decisive factor in accidents on the road [1, 2] and in several other contexts, e. g. in air traffic control. Its detection can be favourable within human-machine communication [3], be this in (serious) games or in any information retrieval system. So far, other information but speech, such as physiological signals [4, 5, 6], or video-based recordings, has been more in the focus of pertinent research. Speech offers, however, some advantages such as non-intrusive recording and robustness against temperature or unfavourable illumination conditions [7]. Moreover, sleepy speech is in itself an interesting paralinguistic phenomenon. The original intention for this paper was two-fold: We first wanted to employ a strict brute-force approach in order to harness every possible information; in addition, we wanted to give it a try and interpret most relevant features. A third objective turned out to be highly interesting, namely the difference between male and female speakers.

2. DATA

We employ the Sleepy Language Corpus (SLC) from the Interspeech 2011 Speaker State Challenge [8, 9]. 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). As we are interested in more realistic speech, 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, [10]), was used by the subjects (self-assessment), and additionally by the three assistants who had supervised the experiments, using all available information (audio, video, context information); 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, struggling 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. A more detailed description of the data is given in [11, 7]. 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*, 2466 turns, 6.6 hours), estimating parameters on the union of the original training and development set (henceforth *TRAIN*, 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 the phoneme recognizer of the Brno University of Technology [12] for segmenting pauses, vowels, consonants, and speaker noise. The 8010 features can be broken into the following subsets:
Formant Features: We use the *snack* toolkit[1] to track the first four formants and their bandwidths; to obtain meaningful results, we restrict the calculation of 12 functionals onto vocalic intervals as segmented by the phoneme recognizer: mean, standard deviation, minimum, maximum, median, quantiles 5%, 25%, 75%, 95%, average

* The authors have received funding from the German Research Council (DFG) under grant agreement KR 3698/4-1. The responsibility lies with the authors.

[1]http://www.speech.kth.se/snack/

absolute local change[2], root average squared local change, and slope of the regression line (appropriately accounting for gaps in between vocalic intervals). This results in $4 \cdot 2 \cdot 12 = 96$ formant features.

MFCC-Features: A more fine-grained, more robust, and less explicit representation of articulators are Mel Frequency Cepstral Coefficients (MFCC). We calculate 13 HTK-compatible MFCC for each speech frame (25 ms, step size 10 ms). Again, we compute a fixed number of features from each coefficient by calculating the 12 functionals described above on all vocalic segments according to the segmentation by the phoneme recognizer ($13 \cdot 12 = 156$ features).

Glottis Features: We perform *glottal speech inversion*, i. e. automatically estimate the free parameters of a glottis model from a given speech sample (described and applied in [14, 15]). The glottis model is a physical mass-spring vocal fold model developed in [16] as described in [17]. Nine parameters determine the physical properties of the model, including the masses, the compliances of the springs, etc. Given the parameter values, the glottis model generates an excitation signal: In short, we first extract the f0 contour using the snack toolkit, and estimate the excitation signal by the LPC residue. Then, the parameters of the model are estimated for each voiced speech frame (25 ms, step size 10 ms) by minimizing the mismatch between observed and predicted f0, and between the spectrum of the observed and predicted excitation signal. The simplex algorithm [18, 19] is used for this non-linear optimization problem; as initialization we use a set of neutral glottis parameters in between typical values for male and female speakers. The result of the glottal speech inversion is a sequence of values for each glottis parameter over voiced intervals. To obtain a fixed number of features, we calculate the same functionals that we used for the formant sequences on each of the parameter sequences, obtaining $9 \cdot 12 = 108$ glottis features.

Prosodic Features: We apply our comprehensive general-purpose prosody module [20] that successfully has been employed for a variety of paralinguistic tasks [21, 22]. The features (raw and adequately transformed/normalized, 79 per segment, using a handful of functionals such as maximum or slope) are based on duration, energy, pitch, and pauses, and can be applied to locally describe arbitrary units of speech such as words or syllables. The sequence of these local features is converted into a fixed-length vector by the four functionals mean, standard deviation, minimum and maximum. Here, we use pseudo-syllables derived from the phoneme recognizer output as units: (1) just the nuclei (i. e. consecutive vowels), (2) nucleus + coda (consecutive vowels plus trailing consecutive consonants), (3) onset + nucleus (leading consonants plus consecutive vowels), and (4) overlapping pseudo-syllables (leading consonants plus consecutive vowels plus trailing consonants). Per utterance, we obtain a total of $79 \cdot 4 \cdot 4 = 1264$ prosodic features.

Rhythm Features: Based on the segmentation of the phoneme recognizer into vocalic and consonantal intervals, we compute: Grabe's raw pairwise variability index rPVI [13] on consonants and vowels plus its (rate-of-speech-) normalized version nPVI. Additionally, we compute variants of Grabe's measures using squared instead of absolute differences (root average squared local change, cf. the formant functionals above). This constitutes 8 features reflecting local variability in durations. Five features reflecting global variability and proportions are given by Ramus' %V (percentage of vocalic intervals) and vocalic and consonantal Deltas (global standard deviations of durations) [23], plus Dellwo's variation coefficient Varco (rate-of-speech-normalized standard deviation of durations) [24].

openSMILE Features: OpenSMILE [25] is a toolkit for computing general-purpose acoustic and prosodic features. It has proven successful for a variety of paralinguistic tasks and was used for providing the baselines in the 2011 challenge [8, 9]. Several low-level descriptors such as loudness, pitch, or energy in spectral bands are modelled by various functionals such as mean, standard deviation or quantiles. We employ the configuration of the Interspeech 2013 Computational Paralinguistics Challenge [26] which results in 6373 features per utterance. No external segmentation is necessary (openSMILE employs its own methods based on energy or voicing).

[2]similar to Grabe's raw pairwise variability index rPVI, [13]

4. EXPERIMENTS AND RESULTS

We do not dichotomize into sleepy/not sleepy as done in the 2011 challenge but work directly with the quasi-continuous KSS sleepiness values. As performance measure, we use Pearson's correlation coefficient r. Spearman's ρ did not differ much, so we skip it. For estimating a speaker's sleepiness score based on the features from an utterance, we apply multiple linear regression, followed by a clipping to the KSS interval $[1, 10]$; ridge regression [27] is used for robust estimation. As a preprocessing step, each feature is transformed individually to standard deviation 1 and mean 0. The scaling ensures that the amount of regularization has a uniform effect, independent of the magnitude of the features. For an easier optimization of the regularization parameter α across different feature sets, the feature vectors are then scaled globally in such a way that the average length (L2-norm) of an instance is 1. The parameters for these transformations are estimated on TRAIN. The metaparameter α is optimized with respect to a 2-fold speaker-independent cross-validation on TRAIN as follows: first, a coarse grid search determines the best α within $\{10^{-6}, 10^{-4}, \ldots, 10^{2}\}$. For speed-up, only 10% of the instances in TRAIN are used in this step. Then, α is refined up to a power of ten with hill climbing, now using all data from TRAIN. Note that it is crucial to optimize α in a *speaker-independent* cross-validation to get a good performance on the unknown speakers in TEST. We use *scikit-learn* [28] for our experiments.

For interpretation, we perform feature selection in some experiments. We use a wrapper approach with a (greedy) sequential forward search: Each time that feature is added which yields the best performance when training *and* testing the regression system with TRAIN. Note that when separating training from testing by using a nested cross-validation within TRAIN for the selection criterion, performance got worse. This can be explained by (1) the larger set of instances used when training and testing with TRAIN together with (2) the fact that overfitting is limited since for each considered feature set, the metaparameter α of the regression system is optimized using cross-validation as described in the previous paragraph.

For a closer look at what happens to individual phonemes in sleepy speech, we segment them with a German speech recognizer. The Kaldi-toolkit [29] is trained with parts of the Verbmobil corpus [30] amounting to 28 hours of conversational speech from 578 speakers. We use an SGMM system with 6000 leaves and 35000 Gaussians, yielding with a pruned trigram a WER of 14.7% on a speaker-independent test set. Since no transcription is available for SLC, we restrict this analysis to the simulated interaction with a driver assistance system where speakers produce scripted utterances. Employing a suitable language model, we account for deviations/omissions and obtain fairly precise phone(eme) boundaries; however, this does not always work for very unclearly spoken words.

4.1. Performance

We first look at the performance of the regression system on TEST when using all available features of a group (see Table 1).

Table 1: *Pearson's correlation r between predicted and reference sleepiness values using different feature groups; best results in bold type. Results are given for training and evaluating on all data ('f+m') and on subsets ('f': females, 'm': males).*

Features	f+m	f	m
Formants	0.27	0.27	0.10
Glottis	0.30	0.26	0.33
MFCC	0.36	0.24	0.52
Prosody	0.34	0.33	0.35
Rhythm	0.20	0.12	0.23
all	0.44	0.41	0.48
openSMILE	0.45	0.42	0.52
all+openSMILE	**0.48**	**0.44**	**0.54**

The performance of the formant, glottis, MFCC, and prosody feature groups (column 'f+m' in Table 1) is around 0.3, only the rhythm features are clearly lower at 0.2. When all features from these groups are used (row 'all'), a correlation of 0.44 is obtained, indicating that complementary information is contained in the sets; note that we have to include all five feature sets to obtain this figure. The openSMILE features yield 0.45, and combining them with all others (row 'all+openSMILE'), we improve the correlation to 0.48. This difference is significant with $p < 0.001$, [3] so we can claim some complementarity between openSMILE and the other features.

We can push performance a bit with an additional preprocessing operation: By scaling the whole feature vector of each instance [32] to length 1, we get a correlation of 0.49. Training gender-dependent regression models, and using an gender classifier[4] for selecting the model applied to each test utterance, we obtain a correlation of 0.50. Together, these two improvements yield a significant difference with $p < 0.03$. However, as both instance scaling and gender dependent models go at the expense of interpretability, these systems are not included in Table 1 and not considered in the interpretation below.

4.2. Gender Differences

In Table 1, performance is clearly higher for male speakers (column 'm') than for female speakers (column 'f') – even though the training set is more than twice as large for females. The MFCC features are the most prominent example: $r = 0.52$ for males vs. $r = 0.24$ for females. Formants, on the other hand, seem to be more suited to predict the sleepiness of females (0.27 vs. 0.10). When all available feature groups including openSMILE are combined (row 'all+openSMILE'), the gap in gender-dependent performance narrows a bit but remains still noticeable (0.44 for females vs. 0.54 for males). This difference could have several reasons: The sleepiness of female speakers may be more difficult to predict due to *more noise* in the features (e. g. caused by a *higher speaker variability*, or due to feature extraction procedures ill-suited for female voices). Another possibility is that female speakers show their sleepiness *to a lesser extent in speech*, resulting in a predicted score that is less related to the reference scores. To tell apart the two effects, let us model the predicted score $\hat{Y}$ as a noisy and biased estimate of the reference score Y with linear regression: $\hat{Y} = \beta_0 + \beta \cdot Y + \epsilon$, with modelling error $\epsilon \sim \mathcal{N}(0, \sigma^2)$, i. e. normally distributed. The lower β (the slope of the regression line), the less sleepiness is shown in the features; the higher σ (the amount of noise), the more non-sleepiness-related variability is present in the features. If female speakers show sleepiness indeed to a lesser extent in speech, one would also expect female speech to sound more *normal*, i. e. less sleepy. This would show up in a negative offset β_0 for females. These measures have to be interpreted carefully because of other factors. There are slightly different mean and standard deviations of sleepiness in TRAIN and TEST; the same holds for female and male subsets. The regression system has an influence, too: Ridge regression tends to produce smaller output ranges in the presence of higher noise due to its regularization. To keep these influences constant, we performed some more experiments on 'all+openSMILE', see Table 2. In three setups, we compare the performance of the same regressor on female and male data. These setups are: train with all (TRAIN='f+m'), train with females (TRAIN='f'), and train with males (TRAIN='m'). We first compare the gender-dependent figures when training with all data (TRAIN='f+m'). Here, noise is equal for females and males ($\sigma = 1.31$), but there is less sleepiness in the female voices (β: 0.28 vs. 0.32). When training the estimator just with female speech (TRAIN='f'), noise is slightly higher for females (σ: 0.81 vs. 0.77) and slopes are equal (0.18). However, we have to consider that this is a *mismatched* evaluation condition for the male speakers: The regressor will adapt to features specifically suited for predicting sleepiness from female voices. The fact that the slopes are equal and the correlation even lower for the female speakers (r: 0.44 vs. 51) speaks in favour of the hypothesis that female voices show sleepiness to a lesser extent. Consistent with that are the results when just training with male speech (TRAIN='m'): Noise is slightly higher for females (σ: 1.42 vs. 1.33), but the slope is much lower for females (β: 0.14 vs. 0.35).

Table 2: *Characteristics of the predicted sleepiness using all features including openSMILE for different gender setups in TRAIN and TEST ('f': females, 'm': males). Symbols are explained in Sect. 4.2.*

TRAIN	TEST	r	β_0	β	σ
f+m	*f*	*0.43*	*-0.15*	*0.28*	*1.31*
	m	0.51	0.43	0.32	1.31
f	*f*	*0.44*	*0.04*	*0.18*	*0.81*
	m	0.51	1.39	0.18	0.77
m	*f*	*0.21*	*-0.84*	*0.14*	*1.42*
	m	0.54	0.67	0.35	1.33

These results suggests that there is a slightly higher non-sleepiness-related variability in female voices, but the main reason for the lower correlation on female data seems to be that sleepiness is shown to a lesser extent in female speech. We reasoned above that this would result in lower sleepiness values for females. That is indeed suggested by the offset values β_0 in Table 2: In two of three cases, β_0 is negative for females, which means that the predicted sleepiness is lower than it should be. In all cases, β_0 is markedly lower for females than for males (β_0: -0.15 vs. 0.43, 0.04 vs. 1.39, -0.84 vs. 0.67, respectively, for the three setups). Thus, there is a slight tendency for female voices to sound less sleepy than they really are (and male voices more sleepy), and a strong tendency for female voices to sound less sleepy than male voices.

4.3. Interpretation of Features

In order to learn which aspects of speech change with sleepiness, and which gender-dependent changes there might be, we inspect how the regressor utilizes the features to produce predictions that are positively correlated with the reference sleepiness values. To make the

[3] One-sided test of difference between dependent correlations [31]

[4] Linear SVM, instance weights for balancing classes, metaparameter optimization as described above, 98% unweighted average recall

interpretation feasible, we first apply a manual feature selection and discard the more complicated and less interpretable functionals and features per group as long as performance does not suffer too much. Then we apply feature selection (see second paragraph in Section 4) to further reduce the number of features to five per studied feature group. Since we use a linear model, interpreting the role of selected features is relatively straightforward: The absolute magnitude of the regression coefficient tells us something about the relative importance (since each feature has been scaled to unit variance, see first paragraph in Section 4); the sign of the coefficient usually tells us whether sleepiness is predicted to rise (positive coefficient) or fall (negative coefficient) with the respective feature. We ignore the intercept in our interpretation.

Space restrictions do not allow us to interpret every set of five most relevant features for every feature group and for both males and females. Moreover, it is well-known that brute-forcing does not necessarily result in most relevant features that easily can be interpreted. We thus restrict ourselves to a few features that can be plausibly interpreted: The more tired, the higher (f1) and the more forward (f2) is the tongue position; f2 bandwidth is wider for tired speech, i. e. it is less pronounced, and might result in more problems for perception [33]. Most relevant MFCC features mirror less rapid articulation (lower rPVI = less local variation) and more (vowel) centralization in sleepy speech (less pronounced extremes). Most relevant prosodic features show for sleepy speech: more disfluencies (here, audible breathing); less variation in energy; less f0-micro variation (jitter); and lower average speech rate.

Figure 1 shows prototypical aspects for sleepy vs. non-sleepy speech. The changes of the mean values are more pronounced for males than for females. The MFCC illustrate that 'sleepy vowels' are produced with higher tongue position and more fronted; when sleepy, speakers reduce articulatory effort by reducing the movements of the articulators; note that overall, resonances for males cluster more together than for females [34]. The same holds for the fricatives which slightly move into the direction of [S]. Thus, sleepy speech is more slurred than non-sleepy speech – not a surprising but at the same time, a reassuring result, corroborating our expectations.

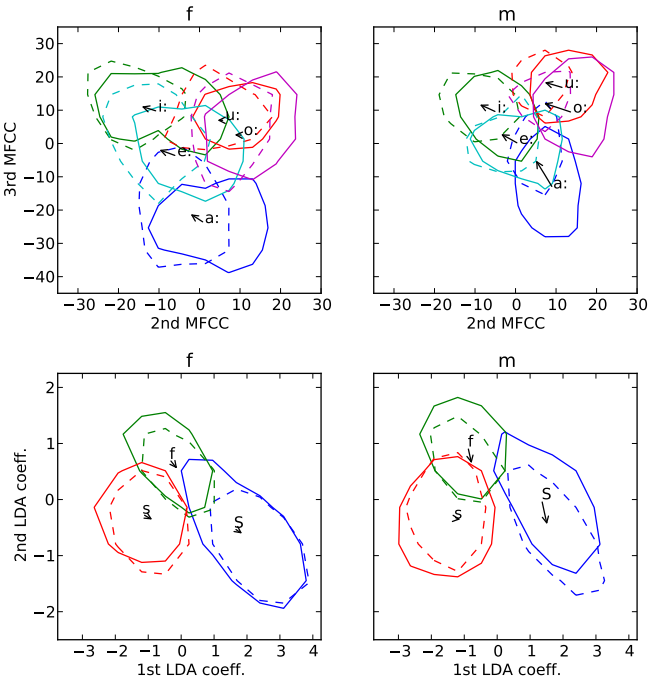

Fig. 1: *MFCC space for the vowels [a:], [e:], [i:], [o:] and [u:] (top) and LDA-projected MFCC space for the consonants [f], [s], and [S] (bottom), for females (left) and males (right). The contour lines include 50% of the probability mass: solid for non-sleepy speech frames (KSS $\leq$ 7.5) and dashed for sleepy speech frames (KSS $>$ 7.5). The arrows indicate how the mean changes between non-sleepy and sleepy speech frames. Used data: simulated interaction with the driver assistance system.*

5. DISCUSSION AND CONCLUDING REMARKS

Overall, sleepiness seems not to be a very clear phenomenon that easily can be predicted when using only audio information. Note that for a dichotomized two-class problem sleepy/not sleepy, the baseline performance in the 2011 challenge was 70.3% unweighted average recall (UA), the best 71.7% [35]. Using our best regression system to feed an SVM classifier for sleepy/non-sleepy (*classification via regression*), we get UA=71.9%[5], similar to what would be expected in the idealized case of normality [36]. Thus, a gender-independent performance of $r = 0.50$ that we obtained for our best system should be a state-of-the-art result. As for different types of features, the brute-force omnibus openSMILE feature set performs markedly better than all other specialized feature sets, and slightly better than a combination thereof ('all' in Table 1). Still, openSMILE together with 'all' yields best performance ('all+openSMILE') – and proves that for performance, brute forcing really pays off.

Feature interpretation is, as often, difficult, as far as highly complex brute-force features are concerned, and meets the expectations, as far as easily interpretable features are concerned: Sleepy speech is more slurred than non-sleepy speech. Most interesting seems to be the question that has been posed in the title of this paper: Are men more sleepy than women – other things being equal, or does it only look like – meaning that the marking of degree of sleepiness for women is less pronounced than for men. The detailed analysis reported in Section 4.2 reveals that the differences between males and females can mainly be traced back to females showing their sleepiness to a lesser degree than men do. In other words: Females tend towards a more canonical pronunciation, (even) when sleepy. [37, p. 130] lists ample evidence and refers to several studies showing that "women tend to produce the rhetorically correct forms of words more often than men do; [...] In contrast, men often reduce or omit vowels and simplify consonant clusters". Obviously, this difference that already [38] has observed persists in sleepy speech as well.

Of course, several caveats hold for the present study, as usual: Our feature vectors are definitely all-encompassing and state-of-the-art; however, there always can be additional information that might be relevant. There are alternatives to our procedures chosen, e. g. other types of feature selection; and of course we have to corroborate the results with other data. At present, we are additionally collecting sleepiness data from more than 100 speakers in an extended experimental design, including transliteration of the data collected.

The marked difference between optimal interlabeller correlations (0.89, reported and discussed in Section 2) and the correlations reported in Table 1 indicates that eventually, multimodal information should be employed for highest performance, if possible. Seen from this angle, this paper contributes to the baselines for audio as a necessary prerequisite.

[5]Not directly comparable, due to our smaller data subset which seems to be a bit easier to process: These 71.9% have to be compared with a UA of 71.5% for the baseline constellation used in the 2011 challenge.

6. REFERENCES

[1] D. Flatley, L. A. Reyner, and J. A. Horne, "Sleep-related crashes on sections of different road types in the UK (1995-2001)," in *Road Safety Research Report*. London: Department for Transport, 2004, vol. 52, pp. 4–132.

[2] T. Horberry, R. Hutchins, and R. Tong, "Motorcycle rider fatigue: A review," in *Road Safety Research Report*. London: Department for Transport, 2008, vol. 78, pp. 4–63.

[3] J. Krajewski, M. Golz, S. Schnieder, T. Schnupp, C. Heinze, and D. Sommer, "Detecting fatigue from steering behaviour applying continuous wavelet transform," in *Proceedings Measuring Behaviour*, vol. 7, 2010, pp. 326–329.

[4] C. Heinze, U. Trutschel, T. Schnupp, D. Sommer, A. Schenka, J. Krajewski, and M. Golz, "Operator fatigue estimation using heart rate measures," in *Proc. World Congress on Medical Physics and Biomedical Engineering*, vol. 25, no. 9, 2009, pp. 930–934.

[5] D. Sommer, M. Golz, T. Schupp, J. Krajewski, U. Trutschel, and D. Edwards, "A measure of strong driver fatigue," in *Proceedings of International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, vol. 4, 2009, pp. 9–15.

[6] R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired?" *Ergonomics*, vol. 51, pp. 982–1010, 2008.

[7] 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.

[8] 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.

[9] 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*, 2013, to appear.

[10] 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.

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

[12] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. of ICASSP*, vol. 1, Toulouse, 2006, p. I.

[13] 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.

[14] P. Beyerlein, A. Cassidy, V. Kholhatkar, E. Lasarcyk, E. Nöth, B. Potard, S. Shum, Y. C. Song, W. Spiegl, G. Stemmer, and P. Xu, "Vocal aging explained by vocal tract modelling: 2008 JHU summer workshop final report," Tech. Rep., 2008.

[15] T. Bocklet, E. Nöth, G. Stemmer, H. Ruzickova, and J. Rusz, "Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis," in *Proc. of ASRU*, Big Island, Hawaii, USA, 2011, pp. 478–483.

[16] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell System Technical Journal*, vol. 51, p. 1233–1268, 1972.

[17] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: The MIT Press, 1998.

[18] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[19] D. Olsson and L. Nelson, "The Nelder-Mead simplex procedure for function minimization," *Technometrics*, vol. 17, no. 1, pp. 45–51, 1975.

[20] 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.

[21] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody – annotation, modelling and evaluation," in *Proc. IS ADEPT, International Symposium on Automatic Detection of Errors in Pronunciation Training*, Stockholm, Sweden, June 6-8 2012, pp. 21–30.

[22] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Tales of tuning – prototyping for automatic classification of emotional user states," in *Proc. Interspeech*, Lisbon, 2005, pp. 489–492.

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

[24] 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.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, New York, NY, USA, 2010, pp. 1459–1462.

[26] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. Interspeech*, Lyon, France, 2013, pp. 148–152.

[27] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, Big Island, Hawaii, USA, 2011.

[30] W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer, 2000.

[31] J. H. Steiger, "Tests for comparing elements of a correlation matrix." *Psychological Bulletin*, vol. 87, no. 2, pp. 245–251, 1980.

[32] A. B. A. Graf, A. J. Smola, and S. Borer, "Classification in a normalized feature space using support vector machines," *IEEE Transactions on Neural Networks*, vol. 14, no. 3, pp. 597–605, 2003.

[33] A. d. Cheveigné, "Formant Bandwidth Affects the Identification of Competing Vowels," in *Proc. of ICPhS*, San Francisco, 1999.

[34] F. Nolan, "Speaker recognition and forensic phonetics," in *The Handbook of Phonetic Sciences*, W. Hardcastle and J. Laver, Eds. Oxford: Blackwell, 1997, pp. 744–767.

[35] D.-Y. Huang, S. S. Ge, and Z. Zhang, "Speaker State Classification Based on Fusion of Asymmetric SIMPLS and Support Vector Machines," in *Proc. of INTERSPEECH*, Florence, Italy, 2011, pp. 3301–3304.

[36] R. Coe, "It's the effect size, stupid," in *Annual Conference of the British Educational Research Association, September 12-14, 2002, University of Exeter, England, Proceedings*, 2002. [Online]. Available: http://www.leeds.ac.uk/ educol/documents/00002182.htm

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

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