Automatic Modelling of Depressed Speech: Relevant Features and Relevance of Gender*

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 Universität München, 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

Depression is an affective disorder characterised by psychomotor retardation; in speech, this shows up in reduction of pitch (variation, range), loudness, and tempo, and in voice qualities different from those of typical modal speech. A similar reduction can be observed in sleepy speech (relaxation). In this paper, we employ a small group of acoustic features modelling prosody and spectrum that have been proven successful in the modelling of sleepy speech, enriched with voice quality features, for the modelling of depressed speech within a regression approach. This knowledge-based approach is complemented by and compared with brute-forcing and automatic feature selection. We further discuss gender differences and the contributions of (groups of) features both for the modelling of depression and across depression and sleepiness.

Index Terms: depression, acoustic features, brute forcing, interpretation, paralinguistics

1. Introduction

Depression is a frequent affective disorder; as a rough estimate, [1] report a one year prevalence of major depression of around 5% for Western European countries. It is characterised, amongst others, by psychomotor retardation, avolition, sadness, maladjusted circadian rhythm (sleep disorder) – all these symptoms showing frequently up in speech in reduction of pitch (variation, range), loudness, and tempo [2], and in voice qualities different from those of typical modal speech. [3] summarise that the "typical speech profile of depression [...] consisted of a triad of reduced stress, monopitch, and monoloudness." [4] report that "the combination of glottal and prosodic features produced better discrimination [of depressed speech] overall than the combination of prosodic and vocal tract features." Not fully in line with these results, [5] point out that detailed spectral features (MFCC) are well suited for detecting depressed speech.

We can expect some similarity between depressed, sleepy, and sad speech: from a functional point of view [6], these states display low arousal and partly negative valence, from a formal point of view, this is mirrored by slowed down and reduced activity in the speech organs: for these speaker states, we might encounter characteristics such as centralisation of vowels, smaller pitch range, lower pitch mean, lower intensity, reduced speech tempo, longer pauses, and atypical voice quality characteristics such as higher breathiness or a tendency towards laryngealised (creaky) speech.

So far, this is the point of view of phonetics: to try and find most important acoustic features. From an engineering point of view, performance has higher priority, i. e., obtaining highest performance in classification or regression. Both points of view have their raison d'être: performance for applications, interpretation for basic research. Here, we want to combine these two different aspects – performance vs. interpretation – in the same way as we addressed acoustic characteristics of sleepy speech in [7, 8] where we used, on the one hand, a very large feature vector and brute forcing, and on the other hand, hand-picked promising features based on the pertinent literature. Moreover, we look into the possibility to transfer the feature vector established for sleepy speech onto depressed speech, while enriching this vector with voice quality information.

2. Data and Annotation

We employ a database with audio recordings of participants in a human-computer interaction experiment, recruited from psychosomatic clinics and universities. The dataset comprises 1122 recordings from 219 German subjects (66 male); mean age 31.5 years, sd 12.9 years, and a range of 18 to 63 years; total duration of all files 29.7 hours. Each speaker filled out a self-assessment questionnaire, amongst others with a score on the Beck Depression Inventory (BDI) scale [9], which ranges from 0 to 63 (severe depression ≥ 29). This highly reliable and standardised questionnaire is the most widely used instrument for measuring the severity of depression. Strong correlations between clinician-rated scales and self-report questionnaires suggest that the two modes of measuring depression may indeed be interchangeable [10]. In our data, mean/sd for recordings of females is 11.2 ± 10.9, of males: 13.1 ± 10.7. The database consists of different tasks: read speech (excerpts of the novel 'Homo Faber' by Max Frisch; the fable 'Der Nordwind und die Sonne' (The North Wind and the Sun)); spontaneous speech (telling a story from the subject's own past describing the best present ever received; telling an imagined story applying the Thematic Apperception Test (TAT), containing, e. g. pictures of a crying person or of a housewife and children who are trying to reach the cookies). The length of the speech tasks is between 5.8 seconds and 5.3 minutes (mean = 1.6 minutes).

We now describe shortly the data from the Sleepy Language Corpus (SLC) from the Interspeech 2011 Speaker State Challenge which we use in a sort of figure-ground manner: we concentrate on depressed speech but, based on the assumed similar-

* 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.

ities of depressed and sleepy speech, discuss the characteristics of depressed speech in relation to those of sleepy speech; more details on SLC are given in [11, 12, 13, 14]. Ninety-nine German speakers (29 male) took part in six partial sleep deprivation studies (mean age 24.9 years, sd 4.2, range 20–52). We use five subsets (read speech: The story of "Der Nordwind und die Sonne" ('the North Wind and the Sun'); commands/requests: simulated driver assistance system commands/requests; 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). The data amount to 7745 recordings with a duration between 0.7 seconds and 3.9 minutes (mean: 9.2 seconds), in total about 20 hours of speech. A well established, standardised subjective sleepiness questionnaire, the Karolinska Sleepiness Scale (KSS, [15]) from 1 (extremely alert) to 10 (extremely sleepy), was used by the subjects (self-assessment) after each recording session, and after all recordings – using all available information (audio/video/context) – by the three assistants who had supervised the experiments. Self-assessment and observer scores are averaged to form the reference sleepiness values (mean/sd: 6.1 ± 2.3 for females, 5.9 ± 2.5 for males).

3. Features

We employ 3805 acoustic features; apart from voice quality features, cf. below, they are described in more detail in [7]; here, we only can give the general idea. For segmenting pauses, vowels, consonants, and speaker noise, we use a phoneme recogniser. Then, pseudo-syllables are derived in four different ways, for instance, by taking nucleus + coda (consecutive vowels plus trailing consecutive consonants). We compute four low-level descriptors on a frame-by-frame[1] 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; additionally, longer-term qualities such as jitter and shimmer are estimated over up to 15 neighbouring syllables [16]. F0 is suitably interpolated, normalised per recorded item, and perceptually transformed. Normalised versions of energy and duration remove phoneme-intrinsic influences. To obtain a fixed number of features per item, we compute twelve functionals characterising statistical and temporal properties of these local descriptors: mean, standard deviation, seven quantiles (minimum, 5%, 25%, median, 75%, 95%, maximum), average absolute local change (similar to Grabe's raw pairwise variability index rPVI [17]), root average squared local change, and slope of the regression line. Depending on the type of descriptor, these functionals are computed across all syllables, across all vocalic frames, or separately across all vocalic and all consonantal frames. Additionally, rhythm features are computed [17, 18, 19].

For modelling voice quality, the features mentioned above already contain jitter and shimmer. To complement these, we add the logarithmic harmonicity-to-noise ratio, and the spectral harmonicity of the openSMILE toolkit [20]. Further, we compute four low-level descriptors on a frame-by-frame-basis with a relatively large frame size of 50 ms suitable for pitch analysis: logarithmic energy, local pitch estimate, a harmonicity measure (ratio of the pitch estimates' autocorrelation to energy), and spectral tilt [21] (logarithmic ratio of first harmonic and F0). The twelve functionals described above are applied separately to the descriptors of all vocalic, and of all consonantal frames.

From this brute-force set, we now present the subset that has been proven suitable for modelling sleepiness [12, 8], adding most promising voice quality features. (A detailed motivation is given in [8].) In the following, due to space restrictions, we just shortly motivate groups of features, apart from the new spectral features which are described in more detail.

Spectral Features (*spec*):

For the spectrum, we choose formants and, as a robust representation of the articulators, MFCC, for modelling centralisation and muscular relaxation (dampening).

Formants: **(1)** The *geometric mean of formants F1–F4* per frame, averaged across vocalic frames. **(2)** The *arithmetic mean of the formant bandwidths of formants F1–F4 per frame*, averaged across vocalic frames. **(3)** The *product of the standard deviations of F1 and F2* across vocalic frames. **(4)** The *average of F1 across vocalic frames*.

MFCC: **(5)** The *average of the second MFCC* across vocalic frames as an estimate of the negative spectral slope. **(6)** The *ratio of the first MFCC averaged across vocalic frames to its average over consonantal frames.* **(7–10)** The *standard deviations of the second and third MFCC computed separately across vocalic and consonantal segments.*

Prosodic Features (*pros*):

The selected prosodic features model relaxation/reduction and loss of tension in several aspects (mono-pitch, mono-loudness, lower range, lower tempo).

Pitch: **(11)** The *average of F0 estimates across vocalic frames.* **(12)** The *standard deviation of F0, normalised to the mean F0*, across vocalic frames. **(13)** The *standard deviation of the syllables' average F0*; here and in the following, we use the 'nucleus + coda' pseudo-syllables. Now we apply our micro-structural prosodic features, where F0 undergoes normalization and perceptual scaling. **(14)** The *syllables' F0 maxima*, averaged. **(15)** The *syllables' F0 minima*, averaged. **(16)** The *F0 slope within syllables*, averaged.

Energy: **(17)** The *standard deviation of the syllables' normalised mean energy.* **(18)** A *medium-term estimate of the relative energy* (computed for energy normalization purposes [16] from up to 15 neighbouring syllables, taking into account phoneme-intrinsic properties), averaged across syllables. **(19)** The *average energy slope within syllables*.

Duration: **(20)** *medium-term estimates of the syllables' relative durations*, averaged. **(21–22)** The *average duration of silent and of filled pauses between syllables.*

Rhythm: **(23)** *Ramus' %V, the percentage of vocalic intervals* [18]. **(24–25)** *Grabe's normalised pairwise variability index nPVI* [17], a rate-of-speech-normalised measure of local durational variability, separately for vocalic and consonantal segments. **(26–27)** *Dellwo's variation coefficient* [19], a measure of global durational variability (rate-of-speech-normalised standard deviations of duration), computed separately for vocalic and consonantal segments.

Voice Quality Features (*vq*):

We chose features modelling breathy, laryngealised [21, 22, 23] (creaky), 'shaky' or otherwise irregular phonation:

(28) *Jitter*, i. e. the average cycle-to-cycle (relative) variation of fundamental frequency, which may rise due to less controlled phonation. **(29)** *Shimmer*, i. e. the average cycle-to-cycle (relative) variation of loudness, which might rise with reduced control, or with increased laryngealisation (creakiness). **(30)** *Raw jitter – the normalised absolute average local change (similar to Grabe's nPVI [17]) of frame-to-frame pitch estimates*. These pitch estimates do not undergo smoothing with context infor-

[1] Frame shift is always 10 ms; frame size depends on the descriptor.

mation; irregular phonation might therefore be reflected in an increased value for this feature. **(31)** *Raw shimmer – the normalised absolute average local change of frame-to-frame loudness* (frame size 50 ms). An increase could be a sign of laryngealisation. **(32)** *logarithmic harmonics-to-noise-ratio (HNR)*, which should decrease with breathy or hoarse phonation. **(33)** *spectral harmonicity*, another measure of harmonic clarity, using the mean of consecutive local min-max differences in the spectrum [20]. **(34)** The *spectral tilt*, averaged over all vocalic frames. A positive tilt should result for laryngealisation, values around zero for modal voices, and negative values for breathy speech [21].

4. Experiments and Results

4.1. Analysis of Feature Groups

For estimating a speaker's BDI or KSS score (cases are recordings, i. e. a whole read or told story) we apply multiple linear regression (for robust estimation, ridge regression [24]). Since the distribution of the BDI scores is skewed, we use Spearman's rank order correlation coefficient ρ between predicted and reference values. (In [8], Pearson's correlation coefficient r was employed; the figures for KSS are thus not identical across the two papers, but very similar.) For the depression data, we evaluate in a 4-fold speaker-independent cross-validation (computing one correlation coefficient for all predictions, which is more conservative than averaging over the correlation coefficient of each fold). For the sleepiness data, we adopt the official (speaker-independent) division of SLC and evaluate on the original test set, using all remaining data for training (i. e. the union of the original training and development set). Features are preprocessed to normalise their scale (details in [7]). The necessary parameters for that, plus the metaparameter α of the ridge regression are estimated on the respective training set (i. e. the training set of the respective fold in the case of depression, or the union of the official SLC training and development set in the case of sleepiness). For the metaparameter, an inner 4-fold[2] speaker-independent cross-validation is used to optimise α up to a power of ten w. r. t. the mean squared prediction error.

For comparison with our knowledge-based feature selection, we also perform a data-driven feature selection, using a so-called wrapper approach, together with a greedy forward search: each time that feature is added which yields the best performance of the regression system in an inner 4-fold (speaker-independent) cross-validation on the respective training set. Due to the metaparameter optimization, this search incurs a triple nested cross-validation for the depression data.

With all features (row 'all (3805)' in Table 1) and evaluating on all speakers, cf. columns 'all', the correlation is 0.44 both for the predicted depression scores (column 'depression') and the sleepiness scores (column 'sleepiness'). Comparing results on female (columns 'f') and male data (columns 'm'), there is a striking difference between the two targets: as already reported in [8], the results for sleepiness are much better for male speakers than for female speakers (0.49 vs. 0.35), even though the majority (73%) of training items is from female speakers. The gender-dependent results for depression do not exhibit such a clear difference (e. g. 0.42 for females vs. 0.40 for males).

Looking at the manually selected features from different groups, voice quality (row 'vq (7)') seems to be most important (all: ρ = 0.38 for depression; 0.36 for sleepiness). It is followed for depression by prosody (all: 0.36) before spectral features (all: 0.29), while it is the other way around for sleepiness (0.28 for spectral features vs. 0.22 for prosody). Regarding gender dependence, Table 1 repeats the findings of [8] for sleepiness: spectral features are more useful for detecting male sleepiness (0.40 vs. 0.20), while prosody is more useful for detecting female sleepiness (0.33 vs. 0.21). For depression, there is only a slight, but similar tendency (spectral features: males 0.31 vs. 0.26; prosody: females 0.35 vs. 0.33). The new voice quality features perform better on male sleepiness (0.40 vs. 0.33); no difference is found in the case of depression (0.36 for both male and female).

Table 1: *Regression performance when predicting depression or sleepiness from different feature groups. All speakers were used in training; Spearman correlation (cross-validated/on test) is reported separately for all, female (f), and male speakers (m). Higher correlation = darker.*

Features (#)	Depression			Sleepiness		
	all	f	m	all	f	m
all (3805)	0.44	0.42	0.40	0.44	0.35	0.49
spec (10)	0.29	0.26	0.31	0.28	0.20	0.40
pros (17)	0.36	0.35	0.33	0.22	0.33	0.21
vq (7)	0.38	0.36	0.36	0.36	0.33	0.40
spec + pros + vq (34)	0.39	0.36	0.39	0.42	0.38	0.40
data-driven sel. (34)	0.36	0.31	0.39	0.37	0.30	0.43

Combining all manually selected features ('spec + pros + vq (34)') improves correlation to 0.39 for depression and 0.42 for sleepiness. Intriguingly, these results come quite close to the performance obtained with the full brute-force set (depression: 0.39 vs. 0.44; sleepiness: 0.42 vs. 0.44). Again, correlations are a bit higher for males (0.39 vs. 0.36 for depression; 0.40 vs. 0.38 for sleepiness).

When selecting the same number of features that we selected manually, i. e. 34 features, in an automatic, data-driven manner ('data-driven sel. (34)'), we get a performance similar to our manual selection (e. g. 0.36 for automatic selection vs. 0.39 for manual in the case of depression, on all). The data-driven search probably fails to outperform our knowledge-driven selection due to the limited amount data, noise, plus the fact that in order to limit computational effort for this combinatorial problem, we used a greedy search.

4.2. Analysis of single Features

We compute Spearman's ρ between the reference values and the individual features of each recorded item. To guarantee strict comparability with the regression results, we use the whole dataset for analysing depression, and the official SLC test set for sleepiness. We show the results for the 'most relevant' of our 34 selected features in Table 2. A feature is defined as 'most relevant' if, across all, females, or males, both for depression and sleepiness, at least one value is ≥ 0.25. By that, we disregard lower values that might be caused by noise (peculiarities of the tasks, speakers, or simply random factors). This arbitrary but meaningful criterion reduces 34 to 11 features. Overall, just by estimating the grey level of the background in Table 2, we can see that the values for depression are higher than those for sleepiness, and there is less variety across genders. For most

[2]The rationale for choosing four folds was to have more than 50% of the data available for training in the inner loop of the double nested cross-validation for the depression data. Four folds yield $\frac{3}{4} \cdot \frac{3}{4} = 56.25\%$ of the data.

Table 2: *Left: Most important manually selected features and their Spearman correlation to the BDI score of the speaker: for all, females (f), and males (m). Right: (also) Spearman correlation for sleepiness KSS. Higher absolute value of correlation = darker.*

Subgroup	**Feature**	**Depression**			**Sleepiness**		
		all	f	m	all	f	m
Formants	(3) product of the standard deviations of F1 and F2	+0.13	+0.11	+0.26	+0.02	−0.04	−0.14
MFCC	(5) average of second MFCC across vocalic frames	−0.09	−0.11	−0.10	−0.24	−0.10	−0.44
	(7) std. deviation of second MFCC across vocalic segments	−0.18	−0.14	−0.25	−0.04	+0.02	−0.23
	(8) std. deviation of third MFCC across vocalic segments	−0.29	−0.24	−0.30	−0.18	−0.18	−0.29
	(10) std. deviation of third MFCC across consonantal seg.	−0.25	−0.24	−0.26	−0.13	−0.09	−0.26
Pitch	(11) average of pitch estimates across vocalic frames	−0.23	−0.24	−0.33	+0.04	−0.26	−0.11
	(13) standard deviation of syllables' average F0	−0.25	−0.21	−0.29	−0.04	−0.07	+0.05
Duration	(20) average syllables' relative durations	+0.17	+0.13	+0.26	+0.24	+0.22	+0.23
Voice Quality	(31) raw shimmer (local change of frame-to-frame loudness)	−0.37	−0.31	−0.46	−0.33	−0.32	−0.26
	(33) spectral harmonicity	−0.33	−0.33	−0.32	+0.01	−0.13	−0.15
	(34) spectral tilt, averaged over vocalic frames	−0.17	−0.18	−0.14	+0.07	+0.09	+0.42

of the features, depression and sleepiness have the same tendencies, although there are a few pronounced exceptions (positive correlation = higher value indicates more severe depression/sleepiness; negative = vice versa).[3]

Feature (3), a measure of the area occupied by the formants, was expected to fall due to centralization. This does not hold for depression (all correlations positive: +0.13, +0.11, +0.26); a possible explanation could be interference by a stronger presence of the first nasal formant. **Feature (5)** can be interpreted as a negative spectral slope; the observed decrease could be due to a stronger lip high-pass, effected by a more closed mouth position, compatible with the expected reduced muscular tension. **Features (7), (8), (10)** model, more robust that (3), the occupied acoustic feature space, and should fall with the expected less crisp pronunciation. The observed correlations for depression and sleepiness (here, strongest for males) match in 17 out of 18 cases (the weak exception is feature (7) for sleepiness of female speakers with $\rho = +0.02$). Level and range of pitch **(features (11), (13))** consistently fall with the BDI score ($\rho \leq -0.21$), perfectly compatible with the expected reduced prosody. For sleepiness, there are again weak exceptions in 2 out of 6 cases. **Feature (20)** rises and thus shows a decreased speech rate. **Features (31), (33), (34)**, mostly negatively correlated, indicate a more breathy phonation for depressive and sleepy speech. One salient exception are sleepy males with a correlation of $+0.42$ for spectral tilt, indicating at least partly laryngealised speech. Similar to the more pronounced centralisation (features (7), (8), (10)), this might be due to males showing their sleepiness more than females do, see Sec. 5.

For nearly all features, depression has a more consistent effect across genders than sleepiness. To quantify this, we take on the one hand the list of correlations of all 34 manually selected features with the target value for females, and for males on the other hand. Then, we compute the correlation between the two lists to assess their similarity; there is much more similarity between the genders for depression than for sleepiness (Spearman's ρ: 0.78 vs. 0.44; Pearson's r: 0.76 vs. 0.40).

[3]Note that for weakly correlated features, 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 occur due to slightly different distributions of feature range and sleepiness score for female vs. male speakers.

5. Discussion and Concluding Remarks

The performance of our relatively small set of selected features does not differ too much from the one obtained by brute forcing with 3805 features. However, with some optimisations, we might expect for brute forcing a gain of some 10% absolute, cf. [7, 8]: leave-one-speaker-out evaluation, instance scaling [25], and gender-dependent models (after a previous automatic gender classification). Moreover, we have to consider that we only model speech and not the other modalities, that the average BDI in our data is 12 ± 11 which represents only minimal or mild depression [9], and that specific features or feature groups such as spectrum or voice quality might be employed in speaker-specific ways, making the modelling noisy across speakers.

In order to detect signs of depression in speech, we transferred and enriched a small, hand-picked feature set originally designed for the detection of sleepiness, which enables a strict comparability of features and procedures across these two states. The competitive performance of this small feature set compared to brute forcing (cf. Table 1) confirms this approach and corroborates the assumption that both states are characterised by a general relaxation/reduction. All feature groups contribute, containing some complementing information.

For sleepy speech, performance for male speakers is much higher than for female speakers. We showed in [7, 8] that this can mainly be attributed to females showing their sleepiness less than males do. Additionally, sleepiness is expressed differently [8]: "male sleepiness is mainly reflected by spectral changes towards less canonical pronunciation [...] whereas female sleepiness primarily implies prosodic changes such as lowered pitch [... This is] in line with our explanation in [7], cf. [26, p. 130] and [27], that women tend towards more canonical speech." Sleepiness is a medium-term state that is 'normal' and not influenced by personality disorders but by the circadian rhythm – although 'atypical' compared to non-sleepy speech. Depression is a long-term state that is 'atypical throughout', and most probably less influenceable by (partly conscious) speaker or speaker group strategies. This might explain the much more systematic tendencies within depression across genders and features, compared to sleepiness. Of course, the caveat has to be made that all this is based on just two samples belonging to the same language (German) and on specific speaking styles.

6. References

[1] E. Paykel, T. Brugha, and T. Fryers, "Size and burden of depressive disorders in Europe," *European Neuropsychopharmacology*, vol. 15, pp. 411–423, 2005.

[2] H. Ellgring and K. R. Scherer, "Vocal Indicators of Mood change in Depression," *Journal of Nonverbal Behavior*, vol. 20, pp. 83–110, 1996.

[3] J. Darby, N. Simmons, and P. Berger, "Speech and Voice Parameters of Depression: A Pilot Study," *J. Comm. Disord.*, vol. 17, pp. 75–85, 1984.

[4] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech," *IEEE Transactions of Biomedical Engineering*, vol. 55, pp. 96–107, 2008.

[5] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," in *Proc. of Interspeech*, Florence, Italy, 2011, pp. 2997–3000.

[6] B. Schuller and A. Batliner, *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*. Chichester, UK: Wiley, 2014.

[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 *Proc. of ICASSP 2014*, 2014, to appear.

[8] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Acoustic-Prosodic Characteristics of Sleepy Speech – between Performance and Interpretation," in *Proc. of Speech Prosody 2014*, to appear. Available: http://www5.cs.fau.de/Forschung/Publikationen/2014/Hoenig14-ACO.pdf.

[9] D. Maust, M. Cristancho, L. Gray, S. Rushing, C. Tjoa, and M. E. Thase, "Chapter 13 – Psychiatric rating scales," in *Handbook of Clinical Neurology*, T. Schlaepfer and C. Nemeroff, Eds. Elsevier, 2012, vol. 106, pp. 227–237.

[10] A. Rush, T. Carmody, H. Ibrahim, M. Trivedi, M. Biggs, K. Shores-Wilson, M. Crismon, M. Toprac, and T. Kashner, "Comparison of self-report and clinician ratings on two inventories of depressive symptomatology," *Psychiatric Services*, vol. 57, pp. 829–837, 2006.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, pp. 383–406, 2001.

[22] A. Batliner, S. Burger, B. Johne, and A. Kießling, "MÜSLI: A Classification Scheme For Laryngealizations," in *Proc. of the ESCA Workshop on Prosody*. Lund: ISCA, 1993, pp. 176–179.

[23] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press, 1980.

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

[25] 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, pp. 597–605, 2003.

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

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