Running head: Non-Linear Dynamics Features for Detecting Sleepiness

Applying Multiple Classifiers and Non-Linear Dynamics Features for Detecting Sleepiness from Speech

Jarek Krajewski[1], Sebastian Schnieder[1], David Sommer[2], Anton Batliner[3], and Björn Schuller[4]

[1] *Experimental Industrial Psychology, University of Wuppertal, Germany*

[2] *Neuro Computer Science and Signal Processing, University of Applied Sciences Schmalkalden, Germany*

[3] *Pattern Recognition, Friedrich-Alexander University Erlangen-Nuremberg, Germany*

[4] *Institute for Human-Machine Communication, Technische Universität München, Germany*

Corresponding Author: Prof. Dr. Jarek Krajewski, Experimental Industrial Psychology, Univ. of Wuppertal (Germany), Gaußstraße 20, 42097 Wuppertal, Tel.: +49 202 439-3945, E-mail: krajewsk@uni-wuppertal.de

Neurocomputing 84 (2012) 65–75

Applying Multiple Classifiers and Non-Linear Dynamics Features for Detecting Sleepiness from Speech

Abstract

Comparing different novel feature sets and classifiers for speech processing based fatigue detectionis is the primary aim of this study. Thus, we conducted a within-subject partial sleep deprivation design (20.00 - 04.00 h, N = 77 participants) and recorded 372 speech samples of sustained vowel phonation. The self-report on the Karolinska Sleepiness Scale (KSS), and an observer report on the KSS, the KSS Observer Scale were applied to determine sleepiness reference values. Feature extraction methods of non-linear dynamics (NLD) provide additional information regarding the dynamics and structure of sleepiness speech. In all, 395 NLD features and the 170 phonetic features which have been computed partially represent so far unknown auditive-perceptual concepts. Several NLD and phonetic features show significant correlations to KSS ratings, e.g., from the NLD features for male speakers the skewness of vector length within reconstructed phase space (*r* = .56), and for female speaker the mean of Cao’s minimum embedding dimensions (*r* = -.39). After a correlation-filter feature subset selection different classification models and ensemble classifiers (by AdaBoost, Bagging) were trained. Bagging procedures turned out to achieve best performance for male and female speakers on the phonetic and the NLD feature set. The best models for the phonetic feature set achieved 78.3% (NaïveBayes) for male and 68.5% (Bagging Bayes Net) for female speaker classification accuracy in detecting sleepiness. The best model for the NLD feature set achieved 77.2% (Bagging Bayes Net) for male and 76.8% (Bagging Bayes Net) for female speakers. Nevertheless, employing the combined phonetic and NLD feature sets provided additional information and thus resulted in an improved highest UA of 79.6% for male (Bayes Net) and 77.1% for female (AdaBoost Nearest Neighbor) speaker.

Keywords: Non-linear Dynamics, Sleepiness, Speech, Bagging, Boosting

Neurocomputing 84 (2012) 65–75

1 Introduction

1.1 Speech Based Sleepiness Measurement

Sleepiness is a crucial factor in a variety of incidents and accidents in road traffic [1-3] and work contexts (e.g., safety sensitive fields such as chemical factories, nuclear power stations, and air traffic control [4,5]). Thus, the prediction and warning of traffic employees against impending critical sleepiness plays an important role in preventing accidents and the resulting human and financial costs. Moreover, the aim to enhance joy of use and comfort within Human-Computer-Interaction (HCI) could also benefit from the automated detection of sleepiness. Knowing the speaker's sleepiness state can contribute to the naturalness and acceptance of HCI by offering empathic feedback about the user's current energetic state. Moreover, understanding the effect of sleepiness on speech production will enhance the robustness of speech and speaker recognition systems.

Hence, a great portion of the literature is based on measuring sleepiness related states. But these mostly electrode- or video-based instruments [6] still do not fulfil the demands of an everyday life measurement system [7]. The major drawbacks are (a) a lack of robustness against environmental and individual-specific variations (e.g., bright light, wearing correction or sun glasses, occlusions, or anatomic variations such as small palpebral fissures) and (b) a lack of comfort and longevity due to electrode sensor application. In contrast to these electrode- or video-based instruments, the utilization of voice communication as an indicator for sleepiness could match the demands of everyday life measurement. Contactless measurements such as voice analysis are non-obtrusive (not interfering with the primary task) and favourable for sleepiness detection since an application of sensors can cause annoyance, additional stress and often impairs working capabilities and mobility demands. In addition, speech is easy to record even under extreme environmental conditions (bright light, high humidity and temperature) even if several sources of noise during driving, such as motor sound, radio, and sidetalk, can lead to difficult recording situations [8,9]. Nevertheless, speech analysis requires merely cheap, durable, and maintenance free sensors and most importantly, it utilizes already existing communication system hardware.

Little empirical research has been done to examine the effect of sleepiness states on voice characteristics. Most studies have analyzed only small phonetic feature sets [10-13] or small feature sets containing only perceptual acoustic features, whereas signal processing based non-linear dynamics features has received little attention [14-16]. Thus, the aim of this study is to apply different phonetic feature sets derived from non-linear dynamics and different classifiers known from biosignal analysis on the detection of critical sleepiness states.

This paper is organized as follows: Section 1.2 introduces the cognitive-physiological mediator-model of sleepiness induced speech changes. Section 2 describes the design of the sleep deprivation study in order to build up a sleepy speaker database. Having provided the results of the sleepiness detection in section 3, the paper closes with a conclusion and a discussion of future work in section 4.

1.2 Sleepiness Induced Non-linear Effects on Speech Production

Speech is modulated by several categories of information, which provide a challenge for detecting specific sleepiness related speech effects [17]: (a) Environmental information: spatial position (place and orientation of source, channel), and background noise (level of vocal effort, Lombard effect); (b) Organic information: biological trait (e.g., height, weight, biological age, age cohort, gender, identity), group and ethnicity membership (e.g., nativeness, regional dialect, race, culture, social class), and pathology (e.g., having a flu, illness, stutter); (c) Expressive information: short term emotion-related states (e.g., stress, intimacy, interest, confidence, uncertainty, deception, politeness, frustration, sarcasm, and pain), temporary states (e.g., sleepiness, medical or alcohol intoxication, health state, mood, depression), interpersonal roles and attitudes (e.g., role in dyads, dominance, submissivity, sympathy, friendship, irony, positive/negative attitude), and long term traits (e.g., personality traits, likeability, temperament); (d) Linguistic information: message content, stressed information, language, and speech style (e.g., whispery, crying, soft, loud, murmuring, clear).

Despite these different factors influencing speech, the temporary states of sleepiness show a distinct pattern of effects on speech. They impair several cognitive components which are relevant for speech production such as early perceptual (visual sensitivity [18]), central („central slowing hypothesis“ [19]), and late motor processing steps („psychomotor slowing“ [20]). These cognitive factors might influence speech production, which can be described as a mental and physical process, which is realized in the following steps: intention to communicate, create an idea, choose suitable linguistic units from memory, form a sentence, generate a sequence of articulatory targets, activate motor programs for the targets, transmit neuromuscular commands to muscles of the respiration and phonation system, move the respective articulators, use the proprioceptive feedback, and radiate the acoustic energy from the mouth and nostrils. Within the stages many feedback loops are involved which help to modify and redirect the process.

Sleepiness related cognitive-physiological changes can influence indirectly voice characteristics according to the following stages of speech production [21]: At the stage of cognitive speech planning a reduced cognitive processing speed might lead to impaired speech planning [22] and impaired neuromuscular motor coordination processes, slowing down the transduction of neuromuscular commands into articulator movement and affecting the feedback of articulator positions. At the stage of muscular actions (see Figure 1), the effects of reduced body temperature and general muscular relaxation might, e.g., lead to a vocal tract softening and thus to a stronger dampening of the signal due to yielding walls. Accordingly, glottal loss and cavity-wall loss for the lower resonant frequencies (formants), and radiation, viscous and heat-conduction loss for the higher formants are expected.

These changes - summarized in the cognitive-physiological mediator model of sleepiness induced speech changes [23] - are based on educated guesses. In spite of the partially vague model predictions referring to sleepiness sensitive acoustic features, this model provides first insights and a theoretical background for the development of acoustic measurements of sleepiness. Nevertheless, little empirical research has been done to examine these processes mediating between sleepiness, speech production, and acoustic features.

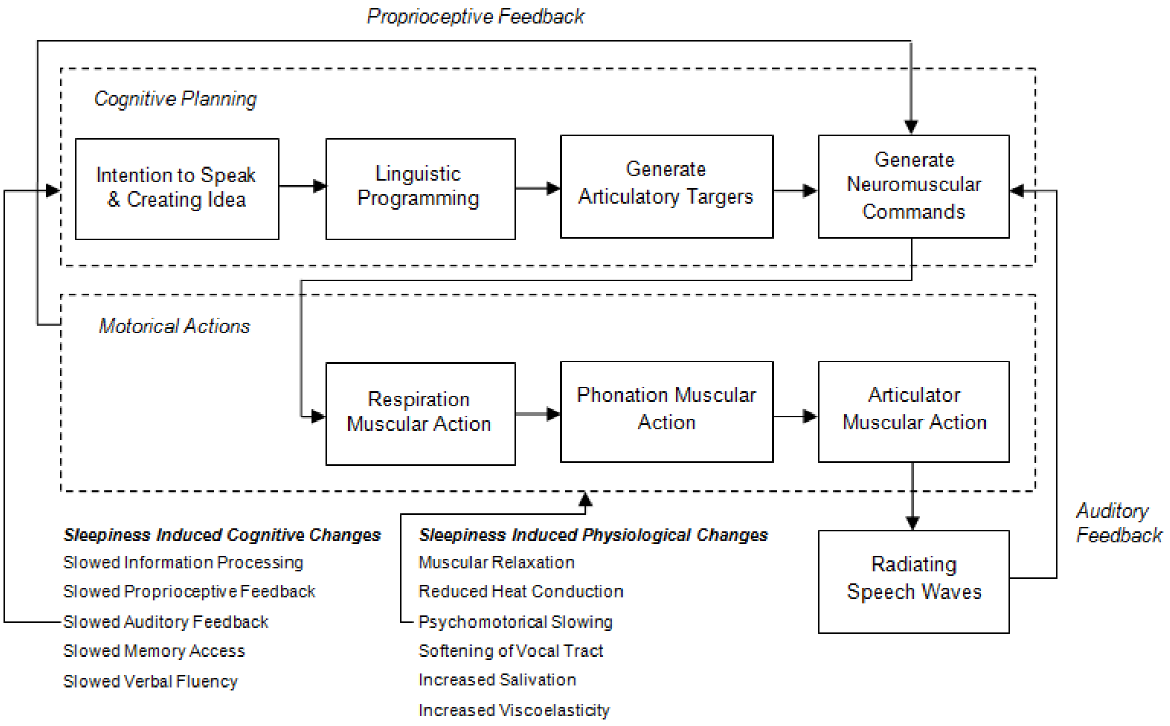
Proprioceptive Feedback
Cognitive Planning
Intention to Speak & Creating Idea
Linguistic Programming
Generate Articulatory Targers
Generate Neuromuscular Commands
Motorical Actions
Respiration Muscular Action
Phonation Muscular Action
Articulator Muscular Action
Auditory Feedback
Radiating Speech Waves
Sleepiness Induced Cognitive Changes
Slowed Information Processing
Slowed Proprioceptive Feedback
Slowed Auditory Feedback
Slowed Memory Access
Slowed Verbal Fluency
Sleepiness Induced Physiological Changes
Muscular Relaxation
Reduced Heat Conduction
Psychomotorical Slowing
Softening of Vocal Tract
Increased Salivation
Increased Viscoelasticity

Fig.1: Speech production process influenced by sleepiness.

An important factor to look at when analyzing sleepiness influenced speech production is generation of non-linear aerodynamic phenomena within the vocal tract. They include non-laminar flow, flow separation in various regions, generation and propagation of vortices and formation of jets rather than well-behaved laminar flow [24-25]. The collapse of laminar flow arises at high Reynolds number. Due to the relevant length and subsonic speed of air flow in the vocal tract, this number is very high. It indicates that the air flow can be expected to be turbulent. The air jet flowing through the vocal tract during speech production includes convoluted paths of rapidly varying velocity, which are highly unstable and oscillate between its walls, attaching or detaching itself, and thereby changing the effective cross-sectional areas and air masses [26].

Several causes are responsible for the generation of these non-linear effects: The vocal folds behave as a vibrating valve, disrupting the constant airflow from the lungs and forming it into regular puffs of air. Modeling approaches which originate in fluid dynamics coupled withelastodynamics of a deformable solid understand this phonation process as non-linear oscillation. Dynamical forcing from the lungs provides the energy needed to overcome dissipation in the vocal fold tissue and vocal tract air. The vocal folds themselves are modeled as elastic tissue with non-linear stress-strain relationship. These non-linear stretching qualities of the vocal folds are based on larynx muscles and cartilage which produces non-linear behavior. Furthermore, vocal tract and vocal folds are coupled when the glottis is open resulting in significant changes in formant characteristics between open and closed glottis cycles. The movement of the vocal folds themselves is modeled by a lumped two mass system connected by springs again with non-linear coupling.

These non-linear phenomena produce turbulent flow while the air jet may be modulated either by the vibration of the walls or by the generated vortices. Several methods based on chaotic dynamics and fractal theory have been suggested to describe these aerodynamic turbulence related phenomena of the speech production system [27-31], including the modeling of the geometrical structures in turbulence (spatial structure, energy cascade) utilizing fractals and multifractals [14,32-35], non-linear oscillator models, and state-space reconstruction [36-38]. This state-space reconstruction is done utilizing the embedding theorem which reconstructs a multidimensional attractor by embedding the scalar signal into a phase space. The embedding allows to reconstruct the geometrical structure of the original attractor of the system which formed the observed speech signal. Moreover, it helps to discover the degree of determinism of an apparently random signal, e.g., by applying measures such as Lyapunov exponents.

However, no empirical research has been done to examine the turbulence effects in speech signals, which might be induced by sleepiness related change of heat conduction within the vocal tract. Previous work associating changes in voice with sleepiness has generally focused only on features derived from speech emotion recognition [16], whereas non-linear dynamics based speech features [27] have received no attention. Thus, the aim of this study is to apply non-linear dynamics (NLD) based features within the field of speech acoustics in order to detect sleepiness.

2 Method

2.1 Procedure, Subject, and Speech Material

A within-subject partial sleep deprivation design (20.00 - 04.00 h, N = 77, 39 female, 38 male participants) containing different speaking tasks was conducted. A well established, standardized subjective sleepiness measure, the self-report on the Karolinska Sleepiness Scale (KSS), was applied to determine the sleepiness reference value (ground truth), which was further used as classification labels. KSS was utilized both by the subjects (self-assessment) as well as by two experimental assistants (observer assessment: assessors were formally trained to apply a standardized set of judging criteria). Scores of the scale used in the present study ranged from 1 to 10, indicating: 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, and cannot stay awake (10). During the night, subjects were confined to the laboratory, had to conduct several human-computer-interaction simulator tasks and were supervised throughout the whole period.

The recording took place in a laboratory room with dampened acoustics using a high-quality, clip-on microphone (sampling rate: 44.1 kHz, quantization: 16 bit). Furthermore, subjects were given sufficient prior practice so that they were not uncomfortable with this procedure. The verbal material chosen for analysis was a three to five second sustained phonation of the vowel /a:/. Total number of speech samples was 372 (188 female and 184 male samples). The recordings (KSS:= mean of the three KSS-ratings; total: M = 5.67; SD = 2.45; females: M = 5.55; SD = 2.15; males: M = 5.89; SD = 2.47) were divided into two classes: not sleepy (‘NSL’, total: 118 samples, 54 female, 64 male) and sleepy (‘SL’, total: 254 samples, 134 female, 120 male) samples with a threshold of 7.5, which was determined by the safety relevant beginning of potential microsleep appearance (about six samples per subject). Participants recorded other verbal material at the same session, however, in this article we only focus on the material described above. Note that we do not pre-select matching pairs, such as instances with a high rater agreement for evaluation. Rather, in line with recent studies in paralinguistic research [39], our goal is to design a system that robustly classifies all available data, as needed for a system operating in a typical daily worktask setting.

Due to data scarcity and a limited amout of parameter optimization feature selection, we did not include a development set, and only distinguished between training (TR) and test set (TE) within a two-fold cross-validation approach. In order to evaluate the final classification accuracy on the TE set, a speaker-dependent hold out validation approach has been chosen. For evaluation of classification tasks, we used the same measures as the INTERSPEECH 2011 Speaker State Challenge: we will use unweighted accuracy (UA) as a measure tailored at imbalanced problemsrememberthat the test set is imbalanced for the two sleepiness classes. For the two-class problem sconsidered in this study, this measure simply reads

$$UA = \frac{\text{(Recall of Class SL ('sleepy') + Recall of Class NSL ('alert'))}}{2}\ .$$

2.2 Feature Extraction

The following feature families within non-linear time series analysis are computed: (a) state space features (360#), and (b) fractal features and entropy features (35#). To extract the non-linear properties of the speech angle signal, we used the time delay embedding method to reconstruct from a scalar time series $\{x\}_{t=1}^{N}$ of N observations (reconstructed phase space, RPS) a vector time series with the embedding dimension *d* and the embedding time lag τvia the transformation

$$x_t \rightarrow (x_t, x_{t-\tau}, x_{t-2\tau}, \dots\ x_{t-(d-1)\tau} \qquad (1)$$

In order to estimate the embedding dimension, a prior knowledge of the fractal dimension of the system would be required. This dilemma is generally solved by employing the false nearest neighbor technique (respectively the first zero of the autocorrelation function for estimating the time lag) or successively embedding in higher dimensions which have been proven successful in former studies.Thus we choose the parameter *d* = 3 and 4, and τ = 1 for the RPS and apply the same embedding procedure to compute a three-dimensional state space by adding delta-regression coefficients (sound pressuret$_0$, Δ sound pressuret$_0$, ΔΔ sound pressuret$_0$).

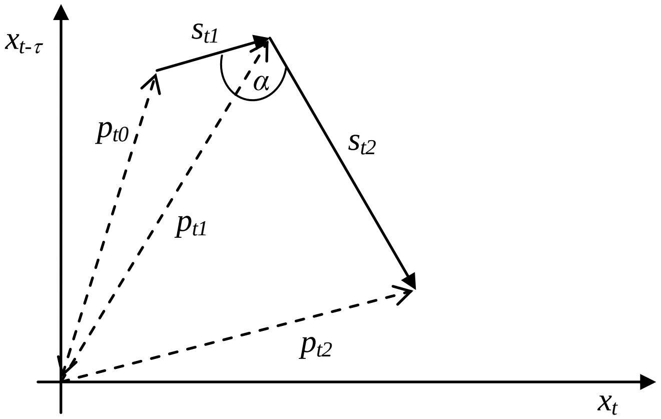

Fig.2. The attractor forming trajectory of three consecutively following vectors p_{t0}, p_{t1},p_{t2}, and geometrical properties of this attractor describing values: angle between consecutive trajectory

parts(α) , the distance to the centroid of the attractor (p_{t0}, p_{t1}, and p_{t2}), and the length of trajectory leg (s_{t1}, s_{t2}).

The geometrical properties of the resulting attractor figures within the state space and the RPS were described by trajectory based descriptor contours,which depict the current angle between consecutive trajectory parts, the distance to the centroid of the attractor, and the length of trajectory leg (see Figure 2 [40]).These trajectory based descriptor contours are joined by their first and second derivatives (velocity and acceleration contours). The temporal information of the contours was captured by computing a time domain functional, which captures temporal information of the contour (see Figure 3). An important advantage of this sequential approach is the improved ability to model the contribution of smaller units and larger units within the structure of a speech sample.The overall 20 time domain functionals used derive from elementary statistics: min, max, range, mean, median, trimmed mean 10%, trimmed mean 25%, 1st, 5th, 10th, 25th, 75th, 90th, 95th, and 99th percentile, interquartil range, mean average deviation, standard deviation, skewness, kurtosis [8,41]. In order to extract features from the RPS, this procedure resulted in an embedding variation for RPS x 3 trajectory based descriptor contours x 3 velocity and acceleration contours x 20 Functionals = 180 RPS features (respectively 180 state space features).

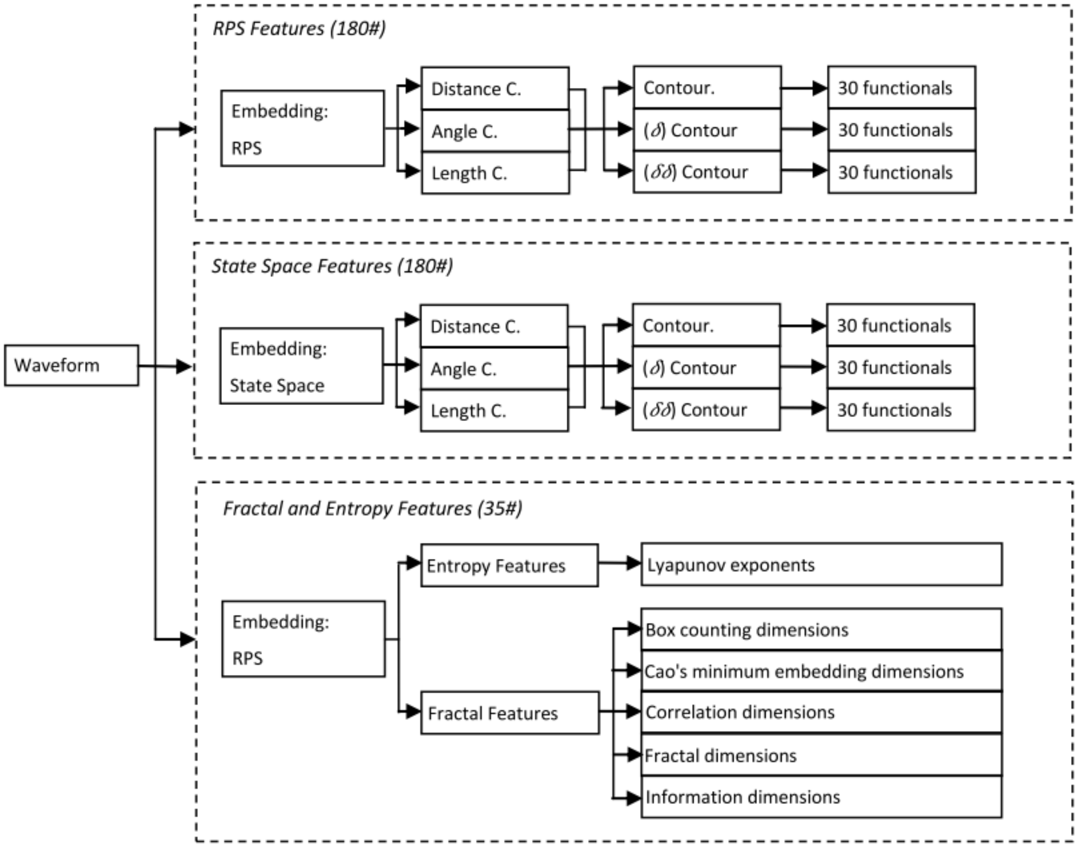

Fig. 3.Process of extracting different classes of NLD features.

The following set of 35 non-linear features derives from the OpenTSTTOOL toolbox available for MATLAB [42] and contains fractal features, which try to quantify self-affinity and underlying complexity of the speech signal and entropy features, assessing regularity/irregularity or randomness of speech signal fluctuations [43]. In detail, the minimum embedding dimension by Cao’s method ([44] max. dimensions = 1,5, delay time = 1,5,10), and the Lyapunov exponents [42], which characterize the system’s “degree of chaos” by estimating the exponential rate of divergence or convergence of nearby orbits on its phase-space, are calculated. Thus, minimum, maximum and mean of the largest Lyapunov coefficient were extracted. Positive Lyapunov exponents indicate divergence of nearby orbits and thus long-term unpredictability. Moreover, state space based prediction using nearest neighbors is applied [45]. The fractal dimension spectrum D(q) is computed using moments of neighbor distances, the boxcounting approach, which computes the boxcounting (capacity) dimension of a time-delay reconstructed time series for dimensions from 1 to D, where D is the dimension of the input vectors using the box counting approach. Capacity dimension D0, information dimension D1, and correlation dimension D2 are extracted by scaling of the correlation sum for time-delay reconstructed timeseries (Grassberger-Proccacia Algorithm). In sum, we computed a total amount of 395 non-linear dynamics features per speech sample.

All phonetic feature extraction of acoustic measurements is done using the Praat speech analysis software [46]. For our study we extracted and computed the following types of features (in total 170), typically used within state-of-the-art speech emotion recognition [16]:

(a) F0 features: The following F0 related features are computed: mean, range, minimum, maximum, 1st, 2nd, and 3rd quartile; (b) Formant position: The analysis of formants indicating the position of the formants 1 to 5 (F1-F5). For each formant mean, range, minimum, maximum, 1st, 2nd, and 3rd quartile as well as differences between mean formant positions are computed; (c) Formant bandwidth: Formant bandwidth displays the peak width of a formant. It is defined as the frequency region in which the amplification differs less than 3 dB from the amplification at the centre frequency. For each formant bandwidth mean, range, minimum, maximum, 1st, 2nd, and 3rd quartile are computed; (d) Intensity: the intensity of a recorded sound can be measured without further modeling of human perception by determining the root mean square amplitude over a short time period. Related to intensity is the popular shimmer feature, which gives an evaluation in percent of the variability of the peak-to-peak amplitude within the analyzed voice sample. It represents the relative period-to-period (very short-term) variability of the peak-to-peak amplitude. The following intensity features are computed: mean, range, minimum, maximum, 1st, 2nd, 3rd quartile and shimmer; (e) MFCC: MFCCs are a widely used variant of the cepstral coefficients that follow the Mel-frequency scale as proposed by psychoacoustics and are the standard features used in Automatic Speech Recognition (ASR). The cepstrum is a measure of the periodicity of a frequency response plot. One of the powerful properties of cepstra is the fact that any periodicities or repeated patterns in a spectrum will be mapped to one specific component in the cepstrum. If a spectrum contains several harmonic series, they will be separated in a way similar to the way the spectrum separates repetitive time patterns in the waveform. Thus, MFCCs offer a compact, decorrelated and accurate high order representation of the speech signal. Computation of the MFCCs includes a conversion of logarithmized Fourier coefficients to the perceptually motivated Mel-scale; (f) Spectral measures: Hammarberg indices were used as another set of coarse measures of spectral properties. Hammarberg, Fritzell, Gauffin, Sundberg, and Wedin [47] showed that differences in voice quality were correlated to differences in the maximum intensity in several frequency bands.

Thus, Hammarberg indices are computed as follows: Hammarberg1 = Maximum spectral amplitude value in frequency band of 400 to 600 Hz minus spectral amplitude of fundamental frequency; Hammarberg2 = Maximum spectral amplitude value in frequency band of 400 to 600 Hz minus spectral amplitude value at 1600 Hz; Hammarberg3 = Maximum spectral amplitude value in frequency band of 400 to 600 Hz minus spectral amplitude value at 5000 Hz; Hammarberg4 = Maximum spectral amplitude value in frequency band of 400 to 600 Hz minus maximum spectral amplitude value above 5000 Hz; the average LTAS spectrum on 6 frequency bands (125-200 Hz, 200-300 Hz, 500-600 Hz, 1000-1600 Hz, 5000-8000 Hz), the proportion of low frequency energy below 500Hz/1000Hz, the slope of spectral energy above 1000 Hz, the Harmonic-to-Noise ratio (HNR), and spectral tilt features ("open quotient", "glottal opening", "skewness of glottal pulse", and "rate of glottal closure") [48].

Machine Learning

Dimensionality Reduction and Classification

The purpose of feature selection is to reduce the dimensionality which otherwise can impede classification performance. The small amount of data given also suggests that longer vectors are not advantageous due to overlearning of data. In this study, we used a relevance maximizing and redundancy minimizing correlation filter approach [49]. This low computational effort demanding technique leads to a compact representation of the feature space. Classifiers typically used within the related field of speech emotion recognition include a broad variety of dynamic algorithms (Hidden Markov Models) and static classifiers [50]. When choosing a classifier within this highly correlated and noisy feature space, several aspects might be of importance such as low memory, low computation time, quick converging, and no suffering from overfitting.

With respect to these requirements (and for transparency and easy reproducibility), we applied the following static classifiers from the popular software RapidMiner, version 4.6 [51] using standard parameter settings: Support Vector Machines ('LibSVM', rbf kernel function, C=0; 'FastLargeMargin' [52], linear kernel, C=1; 'W-SMO', Sequential Minimal Optimization), Logistic Regressions ('MyKLR', dot kernel, C=1; 'LogisticRegression'), Multilayer Perceptrons ('NeuralNetImproved', 2 hidden layer, 5 nodes each; 'W-MultilayerPerceptron', 2 hidden sigmoid layer, 5 nodes each; ‘Perceptron’), k-Nearest Neighbors (‘NearestNeighbors’; k = 1, 5), Decision Trees (‘DecisionTree’, C4.5; ‘RandomForest’, 200 trees), Naive Bayes (‘NaiveBayes’; ‘W-DMNB’, Discriminative Multinominal Naive Bayes), Rule Learner (‘RuleLearner’), Logistic Base (‘LogisticBase’), and Linear Discriminant Analysis (‘Linear Discriminant Analysis’).

Meta-Classification. Various approaches have been suggested to build ensembles of classifiers including the application of different (a) subsets of training data with a single learning method, (b) training parameters with a single training method, and (c) learning methods. Experiments on several benchmark data sets and real world data sets showed improved classification results when using these techniques [53]. In this paper we particularly focus on the two ensembles techniques of Bagging and Boosting [54,55].

Generally speaking, Bagging and Boosting do not try to design learning algorithms which are accurate over the entire feature space: instead they work best for weak learning algorithms fitting in subsamples. They show highest gain for weak classifiers, but have also shown beneficial for strong ones such as SVM or C4.5 trees (leading to Random Forests). The key principle of the bootstrapping and aggregating technique Bagging is to use bootstrap re-sampling to generate multiple versions of a classifier. Bootstrapping is based on random sampling with replacement. Thus, taking a random selection with replacement of the training data can get less misleading training objects (‘outlier’). Therefore, the resulting classifiers may be sharper than those obtained on the training sets with outliers. The second ensemble technique –Boosting - works by repeatedly running a learning algorithm on various distributions over the training data, then combining the classifiers. In contrast to Bagging, where training sets and classifiers are obtained randomly and independently from the previous step, training data is obtained sequentially and deterministic in the Boosting algorithm, reweighting incorrectly classified objects in a new modified training set. Boosting algorithms have also been applied in various research fields, such as natural language processing. In order to determine the added-value of Bagging (sample ratio = 0.9, iterations = 10) and Boosting (iterations = 10) in this application field, we applied these techniques on several commonly applied base-classifiers from the ones named above.

4 Results

Relevance of single features. Reconstructed phase spaces (RPS) figures provide a first insight into possible non-linear dynamics features sensitive to sleepiness. As we can infer from the distances between the trajectories, the sleepy speech attractor figures are less blurred than the alert ones for male and female speaker. As can be seen in the Figures 4 and 5, the standard deviation of the change in the angles is increased in sleepy speakers for both males and females. The mean length and mean angle of an attractor arm are increased for the sleepy speakers.

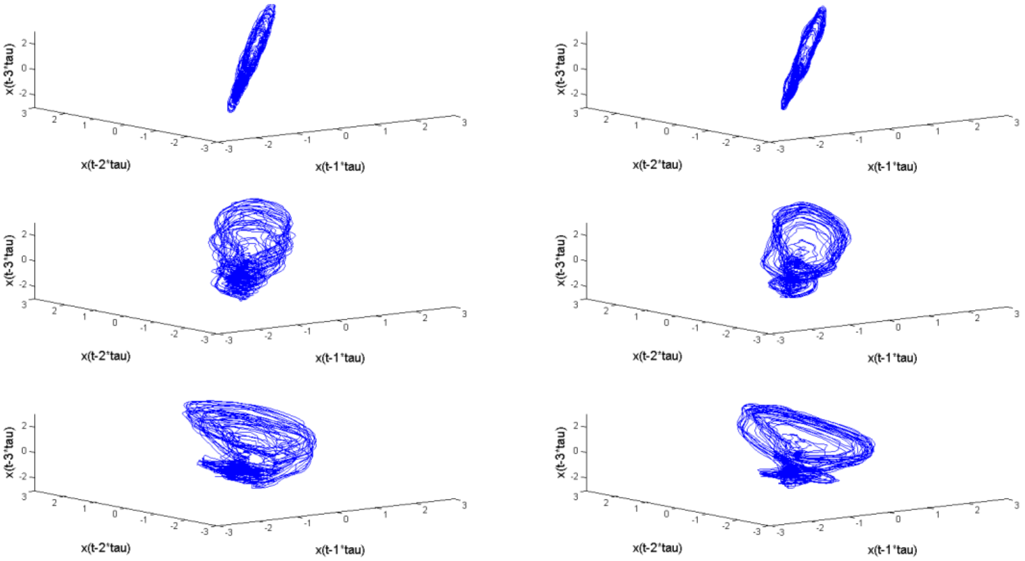

Fig.4. Reconstructed phase spaces (d = 3; τ = 1,5,10) for an alert (NSL, left) and sleepy (SL, right) speech sample /a:/ of a male speaker.

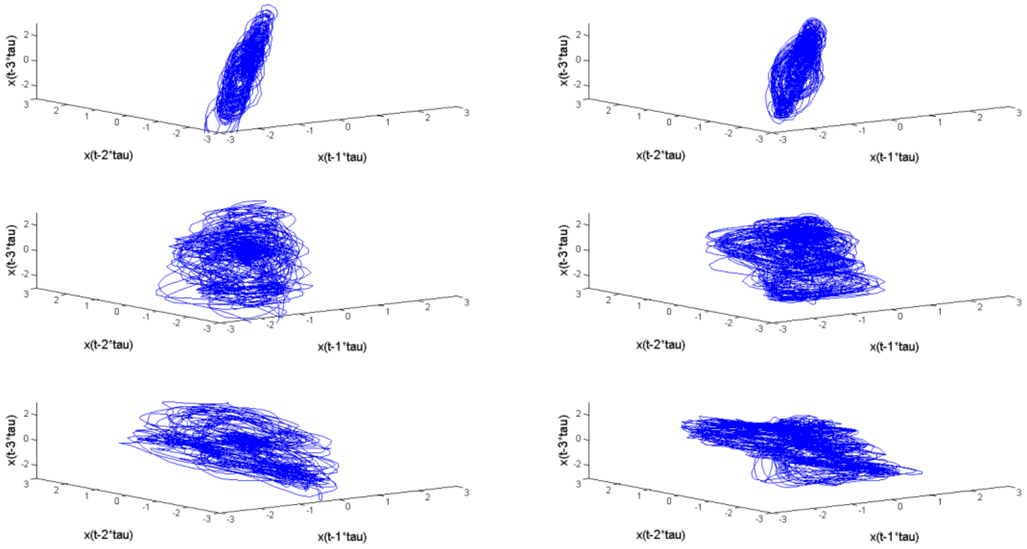

Fig.5. Reconstructed phase spaces (d = 3; τ = 1,5,10) for an alert (NSL, left) and sleepy (SL, right) speech sample /a:/ of a female speaker.

For the NLD feature set, the following features show the highest correlations to KSS rating in males (see Table 1):skewness of vector length within reconstructed phase space = (τ = 1, d = 3), *r* = .56, maximum of Cao's minimum embedding dimensions (τ = 1, dmax = 5), *r* = -.52, and mean of Cao's minimum embedding dimensions (τ = 1, dmax = 5), *r* = -.52. For females, features showing highest correlations to sleepiness are mean of Cao's minimum embedding dimensions (τ = 1, dmax = 5), *r* = -.39, 1st percentile of attractor length within reconstructed phase space (τ = 1, d = 4), *r* = .35, and maximum of Cao's minimum embedding dimensions (τ = 1, dmax = 5), *r* = -.34. These results demonstrate a reduced dimensionality and complexity of a sleepy vowel phonation, as it probably could be achieved by increased regularity and decreased randomness of the speech signal due to changed phonation or turbulence phenomena.

For male speakers the following phonetic features show the highest correlations to KSS ratings (see Table 2): Mel-frequency cepstrum coefficient 1, *r* = -.45, slope of long term average spectrum, *r* = -.39, and Mel-frequency cepstrum coefficient 8, *r* = -.39. For female speakers the features that show the highest correlations to KSS ratings are Hammarberg s3, *r* = .31, Mel-frequency cepstrum coefficient 8, *r* = .31, and Mel-frequency cepstrum coefficient 7, *r* = -.29.

These features are related to sleepiness induced change of voice quality, which represents a shift in vocal fold related phonation processes.

Table 1: Correlation of top 10 NLD features and KSS reference values separated in male and female participants; Critical values for significance tests (*p* < .05) are for male and female *r* = .12.

NLD feature (male)	*r*	NLD feature (female)	*r*
Skewness of vector length within reconstructed phase space = (τ = 1, d = 3)	.56	Mean of Cao's minimum embedding dimensions (τ = 1, dmax = 5)	-.39
Max of Cao's minimum embedding dimensions (τ = 1, dmax = 5)	-.54	1st percentile of attractor length within reconstructed phase space (τ = 1, d = 4)	.35
Trimmed mean of attractor angle within reconstructed phase space (τ = 1, d = 4)	.52	Max of Cao's minimum embedding dimensions (τ = 1, dmax = 5)	-.34
Mean of Cao's minimum embedding dimensions (τ = 1, dmax = 5)	-.52	1st percentile of attractor length within state space	.33
Skewness of attractor angle within reconstructed phase space (τ = 1, d = 4)	-.52	Meanof attractor angle within reconstructed phase space (τ = 1, d = 4)	.32
Max of Cao's minimum embedding dimensions (τ = 1, dmax = 5)	-.51	Mean absolute deviation of delta attractor angle within reconstructed phase space (τ = 1, d = 3)	.31
95th percentil of delta attractor angle within state space	.49	25th percentile of delta attractor length within state space	-.31
5th percentile of delta attractor angle within state space (τ = 1, d = 3)	-.49	Interquartile range of delta attractor length within state space	.30
Standard deviation of delta attractor angle within reconstructed phase space (τ = 1, d = 3)	.48	25th percentile of delta attractor angle within reconstructed phase space (τ = 1, d = 3)	-.30
Standard deviation of delta attractor angle withinstate space	.48	Skewness of attractor angle within reconstructed phase space (τ = 1, d = 3)	-.30

Table 2: Correlation of top 10 phonetic features and KSS reference values separated into male and female participants; Critical values for significance tests (*p* < .05) are for male and female *r* = .12.

Phonetic feature (male)	*r*	Phonetic feature (female)	*r*
Mel-frequency cepstrum coefficient 1	-.45	Hammarbergs3	.31
Slope of long term average spectrum	-.39	Mel-frequency cepstrum coefficient 8	.31
Mel-frequency cepstrum coefficient 8	-.39	Mel-frequency cepstrum coefficient 7	-.29
Percentile 25formant 2bandwidth	.36	Hammarbergs4	.28
ltas_minfreq	.36	Percentile 75 formant 3bandwidth	-.25
Delta Mel-frequency cepstrum coefficient 1	-.36	Percentile 50 formant 3bandwidth	-.25
Minimum of long term average spectrum	.35	Standard deviation of power spectral density	-.23
Range formant 5 Position	-.34	Center of gravity of spectrum	-.18
(PSD frequency band > 1000 Hz) / (PSD frequency band < 1000 Hz)	-.33	Skewness of glottal puls	-.17
Standard deviation of offormand 3position	-.31	Rate of glottal closure	-.17

Overall Classification Results. In order to determine the multivariate classification performance, different classifiers were applied on the phonetic and non-linear dynamics feature sets. The unweighted accuracy (UA) of the different classifiers in male and female speakers for the two class detection problems are listed in Table 3 and Table 4. For the NLD feature set the Bagging Bayes Net classifier achieved the highest UA for male (UA = 77.2%) and the Bagging Bayes Net classifier for female speakers (UA = 76.8%). For the phonetic feature set the NaïveBayes classifier achieved the highest UA for male (UA = 78.3%) and the Bagging Bayes Net classifier for female speakers (UA = 68.5%).The ensemble results are depicted in Table 3 and Table 4. For the NLD feature set, applying just base classifiers would result in an average UA of 66.3% for male and 61.6% for female speakers (Bagging, 67.5% for males and 65.7% for females; AdaBoost, 68.2% for males and 63.5% for females). Within the ensemble classification schemes, the AdaBoost (for males) and the Bagging (for females) algorithms achieved the highest average UA resulting in a slight average improvement over the average single classifier (cf. Table 3).

For the phonetic feature set, applying just base classifiers results in an average UA of 70.6% for male and 60.2% for female speakers (Bagging, 70.1% for males and 60.1% for females; AdaBoost, 69.9% for males and 61.3% for females). Again, the AdaBoost algorithm achieved a slight average improvement over the average single classifier for females (cf. Table 4).

Table 3: Unweighted accuracy rate (UA) in % of several classifiers on the test set using male (m) and female specific models (f) on the NLD feature set: Fast Large Margin = FLM; Sequential Min. Optimization = SMO; Logistic Regression = LR; NeuralNetImproved = NNI; W-Multilayer Perceptron = MLP; Radial Basis Function Network = RBF; 1-NearestNeighbor = 1-NN; 5-NearestNeighbor = 5NN; Decision Tree = DT; RandomForest = RF; NaïveBayes = NB; BayesNet = BN; W-MDNB = MDNB; RuleLearner = RL; LogisticBase = LB; Linear Discriminant Analysis = LDA; Critical values for one-tailed chi-square significance test (*p* < .05) are for male 56.6% and for female: 56.4%.

	Single Classifier						**Bagged Classifier**						**AdaBoosted Classifier**					
	m			**f**			**m**			**f**			**m**			**f**		
	N	**P+N**	**Δ**	**N**	**P+N**	**Δ**	**N**	**P+N**	**Δ**	**N**	**P+N**	**Δ**	**N**	**P+N**	**Δ**	**N**	**P+N**	**Δ**
LibSVM	70.4	74.0	3.6	60.2	61.0	0.8	68.3	76.1	7.8	*76.4*	53.7	-22.7	*73.6*	75.2	1.6	56.3	57.3	1.0
FLM	71.7	60.5	-11.2	65.9	64.2	-1.7	69.5	70.7	1.2	66.4	59.1	-7.3	69.5	70.7	1.2	59.7	67.9	8.2
SMO	72.7	74.5	1.8	60.6	63.6	3.0	*75.2*	75.7	0.5	75.6	59.1	-16.5	72.7	74.3	1.6	61.5	66.8	5.3
MyKLR	64.7	70.4	5.7	58.1	59.3	1.2	61.5	72.0	10.5	61.9	64.0	2.1	60.7	70.8	10.1	62.6	61.8	-0.8
LR	55.7	63.7	8.0	58.9	68.6	9.7	64.3	66.6	2.3	58.7	66.9	8.2	66.6	64.1	-2.5	62.7	66.5	3.8
NNI	64.3	70.8	6.5	59.6	67.0	7.4	64.6	67.9	3.3	68.2	70.3	2.1	69.1	68.4	-0.7	66.7	73.2	6.5
MLP	67.7	68.7	1.0	62.8	70.5	7.7	69.1	65.9	-3.2	62.5	66.3	3.8	68.7	74.4	5.7	60.4	69.4	9.0
PCT	57.3	69.9	12.6	56.0	70.8	14.8	53.9	74.4	20.5	57.2	68.1	10.9	53.0	72.0	19.0	56.1	65.0	8.9
1-NN	62.6	65.5	2.9	59.3	59.2	-0.1	68.3	69.1	0.8	62.3	65.8	3.5	63.8	65.0	1.2	62.3	65.1	2.8
5-NN	60.6	71.9	11.3	63.1	63.8	0.7	67.9	66.3	-1.6	63.6	63.8	0.2	72.3	74.4	2.1	67.9	77.1	9.2
DT	65.1	69.4	4.3	59.0	51.8	-7.2	65.0	65.8	0.8	69.9	68.2	-1.7	65.5	71.5	6.0	63.9	56.6	-7.3
RF	69.2	73.2	4.0	55.6	63.6	8.0	69.2	72.0	2.8	68.9	70.3	1.4	71.4	76.5	5.1	*69.9*	69.4	-0.5
NB	67.6	77.6	10.0	*72.2*	67.2	-5.0	68.4	74.7	6.3	67.1	70.8	3.7	68.4	71.1	2.7	68.7	67.3	-1.4
BN	72.3	79.6	7.3	*70.0*	65.6	-4.4	77.2	78.8	1.6	76.8	68.6	-8.2	75.6	78.4	2.8	68.3	70.8	2.5
MDNB	*73.1*	72.4	-0.7	65.8	63.2	-2.6	73.1	77.2	4.1	62.5	67.8	5.3	71.5	69.9	-1.6	67.4	63.5	-3.9
RL	67.9	59.2	-8.7	62.7	65.2	2.5	64.3	72.4	8.1	58.7	63.7	5.0	71.2	72.1	0.9	56.6	63.7	7.1
LB	71.6	70.8	-0.8	62.6	57.5	-5.1	70.3	65.9	-4.4	61.3	68.1	6.8	67.5	70.8	3.3	66.6	61.2	-5.4
LDA	64.7	70.4	5.7	64.7	60.3	-4.4	64.6	43.0	-21.6	65.4	53.5	-11.9	67.1	69.6	2.5	64.5	53.5	-11.0
Average Classifier	**66.3**	**70.1**	**3.5**	**61.6**	**63.5**	**1.4**	**67.5**	**69.7**	**2.2**	**65.7**	**64.9**	**-0.9**	**68.2**	**71.6**	**3.4**	**63.5**	**65.3**	**1.9**

Table 4: Unweighted accuracy rate (UA) in % of several classifiers on the test set using male (m) and female specific models(f) on the phonetic feature set: Fast Large Margin = FLM; Sequential Min. Optimization = SMO; Logistic Regression = LR; NeuralNetImproved = NNI; W-Multilayer Perceptron = MLP; Radial Basis Function = RBF; Network; 1-NearestNeighbor = 1-NN; 5-NearestNeighbor = 5NN; Decision Tree = DT; RandomForest = RF; NaïveBayes = NB; BayesNet = BN; W-MDNB = MDNB; RuleLearner = RL; LogisticBase = LB; Linear Discriminant Analysis = LDA); Critical values for one-tailed chi-square significance test (*p* < .05) are for male 56.6% and for female: 56.4%.

	Single Classifier						Bagged Classifier						AdaBoosted Classifier					
	m			f			m			f			m			f		
	P	P+N	Δ	P	P+N	Δ	P	P+N	Δ	P	P+N	Δ	P	P+N	Δ	P	P+N	Δ
LibSVM	72.4	74.0	1.6	49.3	61.0	11.7	74.5	76.1	1.6	50.0	53.7	3.7	72.4	75.2	2.8	50.0	57.3	7.3
FLM	67.1	60.5	-6.6	60.6	64.2	3.6	71.1	70.7	-0.4	61.3	59.1	-2.2	69.9	70.7	0.8	64.2	67.9	3.7
SMO	71.6	74.5	2.9	51.5	63.6	12.1	71.2	75.7	4.5	51.1	59.1	8.0	70.3	74.3	4.0	58.3	66.8	8.5
MyKLR	68.0	70.4	2.4	57.9	59.3	1.4	69.6	72.0	2.4	56.3	64.0	7.7	66.7	70.8	4.1	56.3	61.8	5.5
LR	68.7	63.7	-5.0	61.5	68.6	7.1	64.6	66.6	2.0	61.8	66.9	5.1	66.6	64.1	-2.5	61.0	66.5	5.5
NNI	67.9	70.8	2.9	65.5	67.0	1.5	71.1	67.9	-3.2	66.0	70.3	4.3	69.6	68.4	-1.2	65.5	73.2	7.7
MLP	65.1	68.7	3.6	67.4	70.5	3.1	72.7	65.9	-6.8	60.9	66.3	5.4	65.4	74.4	9.0	65.7	69.4	3.7
RBF	77.2	69.9	-7.3	60.1	70.8	10.7	75.6	74.4	-1.2	64.6	68.1	3.5	68.7	72.0	3.3	66.0	65.0	-1.0
1-NN	67.5	65.5	-2.0	58.2	59.2	1.0	65.1	69.1	4.0	61.0	65.8	4.8	68.7	65.0	-3.7	60.3	65.1	4.8
5-NN	75.2	71.9	-3.3	67.1	63.8	-3.3	73.6	66.3	-7.3	67.4	63.8	-3.6	74.0	74.4	0.4	65.6	77.1	11.5
DT	64.6	69.4	4.8	52.2	51.8	-0.4	68.0	65.8	-2.2	51.5	68.2	16.7	65.1	71.5	6.4	50.2	56.6	6.4
RF	74.9	73.2	-1.7	63.8	63.6	-0.2	75.3	72.0	-3.3	67.8	70.3	2.5	75.3	76.5	1.2	67.3	69.4	2.1
NB	78.3	77.6	-0.7	59.9	67.2	7.3	75.1	74.7	-0.4	61.6	70.8	9.2	75.5	71.1	-4.4	67.2	67.3	0.1
BN	71.9	79.6	7.7	70.5	65.6	-4.9	78.0	78.8	0.8	68.5	68.6	0.1	78.0	78.4	0.4	68.5	70.8	2.3
MDNB	76.8	72.4	-4.4	58.6	63.2	4.6	74.0	77.2	3.2	62.7	67.8	5.1	67.9	69.9	2.0	60.7	63.5	2.8
RL	64.0	59.2	-4.8	63.5	65.2	1.7	68.0	72.4	4.4	64.4	63.7	-0.7	72.8	72.1	-0.7	61.6	63.7	2.1
LB	74.4	70.8	-3.6	57.1	57.5	0.4	71.5	65.9	-5.6	53.7	68.1	14.4	70.3	70.8	0.5	53.7	61.2	7.5
LDA	65.8	70.4	4.6	58.1	60.3	2.2	42.5	43.0	0.5	50.6	53.5	2.9	61.5	69.6	8.1	61.0	53.5	-7.5
Average Classifier	**70.6**	**70.1**	**-0.5**	**60.2**	**63.5**	**3.3**	**70.1**	**69.7**	**-0.4**	**60.1**	**64.9**	**4.8**	**69.9**	**71.6**	**1.7**	**61.3**	**65.3**	**4.1**

5 Discussion

The aim of this study was to evaluate a non-linear dynamics (NLD) feature set and different classifier and ensemble classiciation models for speech based sleepiness detection. Advantages of this ambulatory monitoring measurement approach are that in many application settings obtaining speech data is objective and non-obtrusive. Furthermore, it allows for multiple measurements over long periods of time.The main findings may be summarized as following.

First, the NLD and the phonetic features, that were extracted from the sustained vowel phonation and subsequently modeled with machine learning algorithms, contain a substantial amount of information about the speaker's sleepiness state. These results are mainly consistent with the predictions of the cognitive-physiological mediator model of sleepiness. Due to several non-linear phenomena producing turbulent air flow applying NLD speech feature might provide additional information regarding the dynamics and structure of sleepy speech. For example, we achieved significant correlations between sleepiness and the NLD feature 'skewness of vector length within reconstructed phase space' of .56 and the 'maximum of Cao's minimum embedding dimension' of -.54, suggesting a lower complexity of sleepiness speech samples. Explanations for this effect might be found in the following effect chain: sleepiness induced decreased body temperature leading to reduced heat conduction within the vocal tract, changed friction between vocal tract walls and air, changed laminar flows, jet streams, and turbulences, which cause a changed complexity of the speech signal.

Second, we found that NLD and phonetic feature sets result in classification rate high above chance level. Applying an NLD feature set on male speakers separately yielded a UA of 77.2% (Bagging Bayes Net), and a UA of 76.8% for female speakers (Bagging Bayes Net) on unseen data but known speakers. The phonetic feature set achieved a UA of 78.3% (NaïveBayes) for male and of 68.5% for female speakers (Bagging Bayes Net). Unexpectedly, the Bagging Bayes Net classifier nearly consistently outperformed other classifiers. Nevertheless, Random Forest, Support Vector Machine and Neural Network classifier again proved to be among the most promising ones.

Despite of some single NLD features showing higher correlation to sleepiness than phonetic features, the amount of information contained within the total phonetic feature set seemed nearly as high as within the NLD feature set. Nevertheless, employing the combined phonetic and NLD feature sets provided additional information and an average added value for male and female speakers. Thus, further research on NLD features might be a promising challenge.

Third, both ensemble techniques improved the maximum classifier performance for the NLD feature set (male: Bagging, +1.2%, AdaBoost, +1.9%; female: Bagging, +4.1%, AdaBoost, +1.9%) and for the phonetic feature set (male: Bagging, -0.5%, AdaBoost, -0.7%; female: Bagging, -0.1%, AdaBoost, +1.1%). As could be expected, large performance differences between single classifiers could be observed which cannot be fully interpreted. However, the advantage of Average Bagged Classifiers and Average AdaBoosted Classifiers over Average Single Classifiers might be large enough to warrant future experiments. Due to the noisy data, Bagging reached higher performance gains than AdaBoost.

Moreover, comparing the classification results (average classifiers, average bagged classifiers, average adaboosted classifiers) of male and female speakers for the feature sets, it can be determined, that male speakers could be classified more accurately as female speakers, which could be explained by a less balanced class distribution for female speaker. In sum, our classification performance is in the same range that has been obtained for comparable tasks, e.g., for emotional user-state classification [9], which are usually based on much larger databases (over 9,000 speech samples). Thus, it seems likely that sleepiness detection could be improved by collecting similar-sized speech databases, containing speech samples from different types of speakers and speaking styles.

*Limitations and future work.*The validity of our results is limited by numerous facts. Several factors might have influenced the results obtained by NLD methods and, consequently, have to be considered, e.g., recording duration, degree of stationarity, and superimposed noise. Additionally, it would seem beneficial if future studies address the following topics: *Feature extraction:* Evolutionary feature generation methods could be used to explore further features as well as enriching the NLD feature set by further features, e.g., fractal (multifractal analysis, power-law correlation, and detrended fluctuation analysis), entropy (approximate entropy/sample entropy, multiscale entropy, and compression entropy), symbolic dynamics measures, and delay-vector-variance [56,43]. In addition, different normalization procedures could be applied such as computing speaker specific baseline corrections not on high-level features but on duration adapted low-level contours. Additionally, hierarchical functionals [8] might help identifying sleepiness sensitive subparts within a speech segment.

Dimensionality reduction and classification: In order to find ideal feature subsets, further supervised filter based subset selection methods (e.g., information gain ratio) or supervised wrapper-based subset selection methods should be applied (e.g., sequential forward floating search, genetic algorithm selection). Another method for reducing the dimensionality of the feature space are unsupervised feature transformations methods (e.g., PCA network, non-linear autoassociative network, multidimensional scaling, independent component analysis, Sammon map, enhanced lipschitz embedding, or self organizing map or supervised feature transformation methods (e.g., LDA). Last but not least, we should combine the different types of features (NLD and phonetic features) into the same classification pass to find out their combined impact as well as mutual interdependencies. To conclude, methods derived from NLD could offer promising insights into sleepiness induced speech changes. They might supply additional information and complement traditional time- and frequency-domain analyses of speech.

References

[1] D. Flatley, L. A. Reyner, J. A. Horne, Sleep-related crashes on sections of different road types in the UK (1995-2001), in: Department for Transport (Ed.), Road Safety Research Report No. 52, London,2004, pp. 4-132.

[2] T. Horberry, R. Hutchins, R. Tong, Motorcycle rider fatigue: A review, in: Department for Transport (Ed.), Road Safety Research Report No. 78, London, 2008, pp. 4-63.

[3] L. Read, Road safety part 1: Alcohol, drugs and fatigue, in: Department for Transport (Ed.), Road Safety Part 1, London, 2006, pp. 1-12.

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

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

[6] D. Sommer, M. Golz, J. Krajewski, Consecutive detection of driver's microsleep events, in: J. Vander Sloten, P. Verdonck, M. Nyssen, J. Haueisen (Eds.), IFMBE Proceedings 22, Springer, Berlin, 2008, pp. 243-247.

[7] M. Golz, D. Sommer, U. Trutschel, B. Sirois, D. Edwards, Evaluation of fatigue monitoring technologies, J. Somnol. 14 (2010) 187-189.

[8] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, G. Rigoll, Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? Proceedings International Conference on Acoustics, Speech, and Signal Processing 33 (2008) 4501–4504.

[9] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, & G. Rigoll, Non-negative matrix factorization as noise-robust feature extractor for speech recognition, Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (2010) 4562–4565.

[10] Y. Harrison, J. A. Horne, Sleep deprivation affects speech, Sleep 20 (1997) 871-877.

[11] P. Shahidi, S. C. Southward, M. Ahmadian, Estimating crew alertness from speech, Proceedings of the American Society of Mechanical Engineers Joint Rail Conference (2010) 51-59.

[12] A. P. Vogel, J. Fletcher, P. Maruff, Acoustic analysis of the effects of sustained wakefulness on speech, Journal of the Acoustical Society of America 128 (2010) 3747-3756.

[13] J. Whitmore, S. Fisher, Speech during sustained operations, Speech Communication 20 (1996)55–70.

[14] S. McLaughlin, P. Maragos, Nonlinear methods for speech analysis and synthesis, in: S. Marshall, G. Sicuranza (Eds.), Advances in Nonlinear Signal and Image Processing, vol. 6, Hindawi Publishing Corporation, 2007.

[15] H. P. Greeley, J. Berg, E. Friets, J. Wilson, G. Greenough, J. Picone, J. Whitmore, T. Nesthus, Sleepiness estimation using voice analysis, Behaviour Research Methods 39 (2007) 610-619.

[16] B. Schuller, A. Batliner, S. Steidl, F. Schiel, J. Krajewski, The Interspeech 2011 speaker state challenge, Proceedings Interspeech 12 (2011) 3201–3204.

[17] H. Traunmüller, Evidence for demodulation in speech perception, Proceedings of the 6th ICSLP, vol III (2000), 790-793.

[18] P. Tassi, N. Pellerin, M. Moessinger, R. Eschenlauer, A. Muzet, Variation of visual detection over the 24-hour period in humans, Journal of Chronobiology International 17 (2000) 795–805.

[19] D. Bratzke, B. Rolke, R. Ulrich, R., M. Peters, Central slowing during the night, Journal of Psychological Science 18 (2007) 456-461.

[20] D. F. Dinges, N. Kribbs, Performing while sleepy: effects of experimentally induced sleepiness, in: T. H. Monk (Ed.), Sleep, Sleepiness and Performance, J. Wiley & Sons, Chichester, 1991, pp. 97-128.

[21] D. O´Shaughnessy, Speech communications: Human and machine, IEEE Press, New York, 2000.

[22] W. J. M. Levelt, A. Roelofs, A. S. Meyer, A theory of lexical access in speech production, Journal ofBehavioral and Brain Sciences 22 (1999)1-75.

[23] J. Krajewski, B. Kröger, Using prosodic and spectral characteristics for sleepiness detection, in: H. van Hamme, R. van Son (Eds.), Interspeech Proceedings, University Antwerp, Antwerp, 2007, pp. 1841-1844.

[24] H. M. Teager, S. M. Teager, Evidence for nonlinear sound production mechanisms in the vocal tract speech production and speech modelling, in: W. J. Hardcastle, A. Marchal (Eds.), Vol. 55 of NATO Advanced Study Institute Series D,Bonas, 1989, pp. 241–261.

[25] T. J. Thomas, A finite element model of fluid flow in the vocal tract, Comput. Speech & Language 1 (1986) 131–151.

[26] J. F. Kaiser, Some observations on vocal tract operation from a fluid flow point of view, in: I. R. Titze, R. C. Scherer (Eds.), Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control, Denver Center for Performing Arts, Denver, CO, 1983, pp. 358–386.

[27] M. Banbrook, S. McLaughlin, & I. Mann, Speech characterization and synthesis by nonlinear methods, IEEE Trans. Speech Audio Processing 7 (1999) 1–17.

[28] D. Dimitriadis, P. Maragos, Robust energy demodulation based on continuous models with application to speech recognition, Proc. Eurospeech-03 (2003).

[29] D. Dimitriadis, N. Katsamanis, P. Maragos, G. Papandreou, V. Pitsikalis, Towards automatic speech recognition in adverse environments, Proceedings of 7th Hellenic European Conference on Research on Computer Mathematics and Its Applications (2005).

[30] P. Maragos, J. F. Kaiser, T. F. Quatieri, Energy separation in signal modulations with application to speech analysis, IEEE Trans. on Signal Processing 47 (1993) 3024–3051.

[31] S. Narayanan, A. Alwan, A nonlinear dynamical systems analysis of fricative consonants, J. Acoust. Soc. Am. 97 (1995) 2511–2524.

[32] Y. Ashkenazy, The use of generalized information dimension in measuring fractal dimension of time series, Physica A, 271 (1999) 427–447.

[33] O. Adeyemi, F. G. Boudreaux-Bartels, Improved accuracy in the singularity spectrum of multifractal chaotic time series, Proc. IEEE, ICASSP-97 (1997).

[34] P. Maragos, A. Potamianos, Fractal dimensions of speech sounds: Computation and application to automatic speech recognition, J. Acoust. Soc. Am. 105 (1999) 1925–1932.

[35] P. Maragos, Fractal aspects of speech signals: Dimension and interpolation, Proc. IEEE , ICASSP-91 1 (1991) 417-420.

[36] G. Kubin, Synthesis and coding of continuous speech with the nonlinear oscillator model, Proc. ICASSP'96 1 (1996) 267-270.

[37] T. F. Quatieri, E. M. Hofstetter, Short-time signal representation by nonlinear difference equations, Proc. IEEE , ICASSP'90 (1990).

[38] B. Townshend, Nonlinear prediction of speech signals, IEEE Trans. Acoust., Speech, Signal Processing (1990).

[39] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, Speech Communication 53 (2011) 1062-1087.

[40] I. Mierswa, K. Morik, Automatic feature extraction for classifying audio data, Machine Learning Journal 58 (2005) 127–149.

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

[42] C. Merkwirth, U. Parlitz, I. Wedekind, D. Engster, W. Lauterborn, OpenTSTOOL User Manual Version 1.2, Drittes Physikalisches Institut, Universität Göttingen, 2009.

[43] A. Voss, S. Schulz, R. Schroeder, M. Baumert, P. Caminal, Methods derived from nonlinear dynamics for analysing heart rate variability, Philos. Transact. A. Math. Phys. Eng. Sci. 367 (2009) 277–296.

[44] L. Cao, Practical method for determining the minimum embedding dimension of a scalar time series, Physcai D 110 (1997) 43-50.

[45] J. McNames, A nearest trajectory strategy for time series prediction, Proc. of the International Workshop on Advanced Black-box Techniques for Nonlinear Modeling(1998) 112-128.

[46] P. Boersma, PRAAT, a system for doing phonetics by computer, Glot International 5 (2001) 341–345.

[47] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, L. Wedin, Perceptual and acoustic correlates of abnormal voice qualities, ActaOtolaryngol. 90 (1980) 441-451.

[48] K. Stevens,Acoustic phonetics, MIT Press, Massachusetts, 2000.

[49] M. A. Hall, Correlation-Based Feature Selection for Machine Learning, PhD thesis, Department of Computer Science, University of Waikato, 1994.

[50] Z. Zeng, M. Pantic,G. I. Roisman, T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 39–58.

[51] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, &T. Euler, YALE: Rapid prototyping for complex data mining tasks, in: T. Eliassi-Rad, L. H. Ungar, M. Craven, D. Gunopulos (Eds.),

Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, 2006, pp. 935-940.

[52] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, C. J. Lin, Liblinear: A library for large linear classification, Journal of Machine Learning Research 9 (2008) 1871-1874.

[53] L. Breiman, Bagging Predictors,Machine Learning 24 (1996) 123-140.

[54] S. B. Kotsianti, D. Kanellopoulos, Combining bagging, boosting and dagging for classification problems, Knowledge based intelligent information and engineering systems, Lecture Notes in Computer Science 4693 (2009) 493-500.

[55] P. Boinee, A. De Angelis, G. L. Foresti, Ensembling Classifiers - An application to image data classification from Cherenkov telescope experiment, IEC 2005 (2005) 394-398.

[56] D. Sommer, Analyse des Mikroschlafs mit Methoden der Computergestützten Intelligenz, Shaker, Aachen, 2009.

Vitae

Jarek Krajewski received his diploma in 2004 and his doctoral degree for his study on Acoustic Sleepiness Detection in 2008, both in psychology and signal processing from Univ. Wuppertal and RWTH Aachen. He is Assistant Professor in Experimental Industrial Psychology since 2009 and vice director of the Center of Interdisciplinary Speech Science at the Univ. Wuppertal. Prof. Krajewski (co-)authored more than 50 publications in peer reviewed books, journals, and conference proceedings in the field of sleepiness detection, and signal processing.

Sebastian Schnieder received his Master of Science degree in 2008 from Middlesex University, London, UK after completing his studies in Applied Psychology. Since 2009, he is a member of the research group around JarekKrajewski, Experimental Industrial Psychology in Wuppertal, Germany. Sebastian is now working towards his PhD focusing on applying pattern recognition and non-linear prediction models to emotional and psychopathological states. Several presentations and workshops on international conferences document the practical relevance of his research topics. He is a graduate member of the British Psychological Society as well as the American Psychological Association.

David Sommer received his diploma from Univ. of Applied Sciences Schmalkalden in 1998 and his doctoral degree from TU Ilmenau in 2009, both in bio-signal processing and pattern recognition. Since 1998, he has been a scientific co-worker at the Department of Computer Science in Schmalaklden and an Associate Lecturer in neural networks and pattern recognition. David is author/co-author of more than 70 publications in peer reviewed books, journals, and conference proceedings in the field of neural networks, evolutionary algorithms, nonlinear signal processing, data fusion and pattern recognition in different areas of applications, such as driver fatigue, posturography, speech processing and sleep physiology. He serves as reviewer for several journals and conferences, and as invited speaker, session and chairman of several international workshops and conferences.

Anton Batliner received his M.A. degree in Scandinavian Languages and his doctoral degree in phonetics in 1978, both at LMU Munich/Germany. He has been a member of the research staff of the Pattern Recognition Lab at FAU Erlangen/Germany since 1997. He is co-editor of one book and author/co-author of more than 200 technical articles, with a current H-index of 30 and more than 3,300 citations. His research interests are all aspects of prosody and paralinguistics in speech processing. Dr. Batliner repeatedly served as Workshop/Session (co-)organiser and guest editor; he is Associated Editor for the IEEE Transactions on Affective Computing.

Björn Schuller received his diploma in 1999 and his doctoral degree in 2006, both in electrical engineering and information technology from TUM in Munich/Germany where he is tenured as Senior Researcher and Lecturer in Pattern Recognition and Speech Processing. From 2009 to 2010 he was with the CNRS-LIMSI in Orsay/France and a visiting scientist in the Imperial College London's Department of Computing in London/UK. Dr. Schuller is a member of the ACM, HUMAINE Association, IEEE and ISCA and (co-)authored more than 200 peer reviewed publications leading to more than 2 300 citations - his current H-index equals 24.

Author notes

Jarek Krajewski, Experimental Industrial Psychology, University of Wuppertal (Germany); Sebastian Schnieder, Experimental Industrial Psychology, University of Wuppertal (Germany); David Sommer, Neuro Computer Science, and Signal Processing, Univ. of Applied Sciences, Schmalkalden (Germany); Anton Batliner, Pattern Recognition, Friedrich-Alexander University of Erlangen-Nürnberg (Germany); Björn Schuller, Institute for Human-Machine Communication, TechnischeUniversitätMünchen (Germany).

The authors thank Aleksandra Andrejeva and Tom Laufenberg for their helpful commentary and assistance on a preliminary version of this manuscript.