# Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach

**Jarek Krajewski**
*University of Wuppertal, Wuppertal, Germany*

**Anton Batliner**
*University of Erlangen-Nürnberg, Erlangen, Germany*

AND

**Martin Golz**
*University of Applied Sciences Schmalkalden, Schmalkalden, Germany*

This article describes a general framework for detecting sleepiness states on the basis of prosody, articulation, and speech-quality-related speech characteristics. The advantages of this automatic real-time approach are that obtaining speech data is nonobstrusive and is free from sensor application and calibration efforts. Different types of acoustic features derived from speech, speaker, and emotion recognition were employed (frame-level-based speech features). Combing these features with high-level contour descriptors, which capture the temporal information of frame-level descriptor contours, results in 45,088 features per speech sample. In general, the measurement process follows the speech-adapted steps of pattern recognition: (1) recording speech, (2) preprocessing, (3) feature computation (using perceptual and signal-processing-related features such as, e.g., fundamental frequency, intensity, pause patterns, formants, and cepstral coefficients), (4) dimensionality reduction, (5) classification, and (6) evaluation. After a correlation-filter-based feature subset selection employed on the feature space in order to find most relevant features, different classification models were trained. The best model—namely, the support-vector machine—achieved 86.1% classification accuracy in predicting sleepiness in a sleep deprivation study (two-class problem, $N = 12$; 01.00–08.00 a.m.).

Sleepiness impairs a series of cognitive abilities, such as early perceptual (visual sensitivity; Tassi, Pellerin, Moessinger, Eschenlauer, & Muzet, 2000), central (central slowing hypothesis; Bratzke, Rolke, Ulrich, & Peters, 2007), and late motor-processing (psychomotor slowing; Dinges & Kribbs, 1991) steps. Furthermore, the decrements in the speed and accuracy of various task performances can be explained by changes in working memory, executive function, supervisory control (Jennings, Monk, & van der Molen, 2003; Nilsson et al., 2005), spatial orientation, situational awareness (see Harwood, Barnett, & Wickens, 1988), mathematical processing, motor task abilities (e.g., manual dexterity, grip strength, tapping, fine motor control; Durmer & Dinges, 2005; Rogers, Dorrian, & Dinges, 2003; Wesensten, Belenky, Thorne, Kautz, & Balkin, 2004), and divergent-thinking capacity (Horne, 1988; Linde & Bergström, 1992).

Due to these impairments, sleepiness is a factor in a variety of incidents and accidents in road traffic (e.g., Flatley, Reyner, & Horne, 2004; Horberry, Hutchins, & Tong, 2008; Read, 2006) and work (e.g., safety sensitive fields, such as chemical factories, nuclear power stations, and air traffic control; Melamed & Oksenberg, 2002; Wright & McGown, 2001) contexts. Accordingly, 21% of the reported incidents mentioned in the Aviation Safety Reporting System (including those involving pilots and air traffic controllers) were related to fatigue. Thus, the prediction of and warning of traffic employees against impending critical sleepiness play an important role in preventing accidents and the resulting human and financial costs.

In addition to the commonly accepted fact of sleep-induced cognitive impairments, previous research lends some support for the mood-disturbing effects of sleepiness (see Engle-Friedman et al., 2003). Drawing from these findings, we assume that in analogy to the sleepiness-induced decrease of performance within the transportation sector, the performance of communication-centered services will also suffer from sleepiness-related impairments. In addition to sleepiness-induced disturbances in human-to-human communication, human–computer interaction (HCI) could also benefit from the detection of and automatic countermeasures to sleepiness. Knowing

**J. Krajewski, krajewsk@uni-wuppertal.de**

the speaker's sleepiness state can contribute to the naturalness of HCI. If the user shows unusual fatigue states, giving feedback about this fact would make the communication more empathic and human-like. This enhanced naturalism might improve the acceptance of these systems. Furthermore, it may result in better comprehensiveness, if the system output is adapted to the user's actual fatigue-impaired attentional and cognitive resources.

Hence, many efforts have been reported in the literature to measure fatigue states (Sommer, Chen, Golz, Trutschel, & Mandic, 2005). These systems have focused mainly on (1) saccade eye movement (Zils, Sprenger, Heide, Born, & Gais, 2005), instability of pupil size (Wilhelm et al., 2001), and eye blinking (Ingre, Åkerstedt, Peters, Anund, & Kecklund, 2006; Schleicher, Galley, Briest, & Galley, 2008); (2) EEG data (Davidson, Jones, & Peiris, 2007; Golz, Sommer, Holzbrecher, & Schnupp, 2007); and (3) behavioral expression data (gross body movement, head movement, mannerism, and facial expression; Vöhringer-Kuhnt, Baumgarten, Karrer, & Briest, 2004) in order to characterize the sleepiness state. Apart from these promising advances in analyzing eye movement and behavioral expression data, there has recently been renewed interest in vocal expression and speech analysis. This fact is promoted mainly by the progress in speech science and the gaining presence of speech in voice-guided HCI. Using voice communication as an indicator of sleepiness would have the following advantages: Obtaining speech data is nonobstrusive, free from sensor application and calibration efforts, robust against extreme environmental conditions (humidity, temperature, and vibrations), and "hands- and eyes-free," and most importantly, speech data are omnipresent in many daily life situations.

Little empirical research has been done to examine the effect of sleepiness states on acoustic voice characteristics. Most studies have analyzed only single features (Harrison & Horne, 1997; Whitmore & Fisher, 1996) or small feature sets containing only perceptual acoustic features, whereas signal-processing-based speech and speaker recognition features (e.g., mel frequency cepstrum coefficients [MFCCs]; see Table 1) have received little attention (Greeley et al., 2007; Nwe, Li, & Dong, 2006). Building an automatic sleepiness detection engine reaching sufficient precisions still remains undone. Thus, the aim of this study is to apply a state-of-the-art speech emotion recognition engine (Batliner et al., 2006; Vlasenko, Schuller, Wendemuth, & Rigoll, 2007) on the detection of critical sleepiness states. Attention is drawn particularly to the computation of a 45,088-feature set using frame-level descriptors (FLDs) and their temporal-information-aggregating functionals (see Figure 2).

The rest of this article is organized as follows. The second section introduces the cognitive-physiological mediator model of sleepiness-induced speech changes. In the third section, the procedures for computing FLDs and functionals are explained. The fourth section describes the design of the sleep deprivation study used for building a sleepy speaker database. Having provided the results of the sleepiness detection in the fifth section, the article closes with a conclusion and a discussion of future work.

## Sleepiness and Speech Changes

Sleepiness-related cognitive-physiological changes—such as decreased muscle tension or reduced body temperature—can indirectly influence voice characteristics according to the following stages of speech production (O'Shaughnessy, 2000).

1. Cognitive speech planning: reduced cognitive processing speed (*central slowing hypothesis*; Bratzke et al., 2007) → impaired speech planning (Levelt, Roelofs, & Meyer, 1999) and impaired neuromuscular motor coordination processes (*psychomotor slowing*; Dinges & Kribbs, 1991) → impaired fine motor control and slowed articulator movement → slackened articulation and slowed speech.

2. Respiration: decreased muscle tension → flat and slow respiration → reduced subglottal pressure → lower fundamental frequency, intensity, articulatory precision, and rate of articulation.

3. Phonation: decreased muscle tension → increased vocal fold elasticity and decreased vocal fold tension; decreased body temperature → changed viscoelasticity of vocal folds → shift in the spectral energy distribution; breathy and lax voice → nonraised larynx → decreased resonance frequencies (formants) positions and broadened formant bandwidth.

4. Articulation/resonance: decreased muscle tension → unconstricted pharynx and softening of vocal tract walls → energy loss of the speech signal → broader formant bandwidth; postural changes → lowered upper body and lowered head → changed vocal tract shape → changed formant position; increased salivation → energy loss; decreased body temperature → reduced heat conduction, changed friction between vocal tract walls and air, changed laminar flows, jet streams, and turbulences → energy loss → shift in the spectral energy distribution, broader formant bandwidth, increase in formant frequencies especially in lower formants.

5. Radiation: decreased orofacial movement, facial expression, and lip spreading (*relaxed open mouth display*; Kienast & Sendlmeier, 2000; Tartter, 1980) → lengthening of the vocal tract → lower first and second formant positions; reduction of articulatory effort → smaller opening degree → slackened articulation → decreased first formant; oropharyngeal relaxation → lowering velum → coupling of nasal cavity → increased nasality → broadened Formant 1 bandwidth, smaller Formant 1 amplitude.

These changes—summarized in the cognitive-physiological mediator model of sleepiness-induced speech changes (Krajewski, 2008)—are based on educated guesses. In spite of the partially vague model predictions referring to sleepiness-sensitive acoustic features, this model provides a first insight into and 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.

## Brute Force Feature Extraction

**Basic acoustic features**. Acoustic features can be divided according to auditive-perceptual concepts in pros-

ody (pitch, intensity, rhythm, pause pattern, and speech rate), articulation (slurred speech, reduction and elision phenomena), and speech quality (breathy, tense, sharp, hoarse, or modal voice). Another distinction can be drawn from using signal-processing categories such as time domain, frequency domain, or state space features. Our approach prefers the fusion of perceptual features with purely signal-processing- and speech-recognition-based features, without any known auditive-perceptual pendants. Typical frame-level-based acoustic features used in emotion speech recognition and audio processing (see Table 1) are fundamental frequency ($f0$), energy, harmonics-to-noise ratio (HNR), formant position and bandwidth ($F1$–$F6$), MFCCs, linear frequency cepstrum coefficients (LFCCs), duration of voiced/unvoiced speech segments, and spectral features derived from the long-term average spectrum (LTAS), such as band-energies, roll-off, centroid or spectral flux.

**FLDs and FLD contour descriptors (functionals).** The trend in speech emotion recognition is toward an exact temporal modeling of the acoustic feature contours (FLDs), resulting in hundreds or even thousands of features used for classification (Batliner et al., 2006; Schuller et al., 2007). This thorough (*brute force*) exploitation of the feature space is achieved by a four-step process (see Figure 1).

Step 1: The speech signal is split into small frames (e.g., 10 msec) and multiplied with smoothening window functions such as hamming or hanning.

Step 2: Signal-processing-, speaker-recognition-, and speech-recognition-based acoustic features are computed for each single frame.

Step 3: The values of each frame-level feature are connected to FLD contours and joined by their first and second derivates (velocity contour = delta FLD contour; acceleration contour = delta delta FLD contour). This

procedure results in speech feature contours such as the fundamental frequency contour or the second derivate of the bandwidth of the Formant 4 contour.

Step 4: The values of the univariate time series (FLD) are projected onto a scalar feature $x$, which captures temporal information of the acoustic contour (FLD). An important advantage of this sequential approach is the improved ability to model the contribution of smaller units (words) and larger units (chunks) within the prosodic structure of an utterance. Frequently used functionals are percentiles (quartiles, quartile ranges, and other percentiles), extremes (min/max value, min/max position, range), distributional functions (number of segments/intervals/reversal points), spectral functionals (discrete cosine transform coefficients), regression functions (intercept, error, regression coefficients), higher statistical moments (standard deviance, skewness, kurtosis, length, and zero-crossing rate), means (arithmetic mean and centroid), and sequential and combinational functionals: a minimum of two functionals has to be applied in either a sequential way (e.g., maximum of regression error) or a combinational way (e.g., ratio of mean of two different FLDs) (Schuller, Wimmer, Mösenlechner, Kern, & Rigoll, 2008).

## Speech-Adapted Pattern Recognition Framework

The computationally demanding feature extraction procedure usually results in a huge number of acoustic features ($>1,000$) and a comparatively small number of samples (recorded speech segments such as phrases). This problem is well known as the curse of dimensionality and can impair the reliability of classification. Thus, the optimization of high-dimensional feature spaces seems a must in view of performance and real-time capability. The

### Table 1
### Basic Acoustic Feature Contours (Frame-Level Descriptors)

| Frame-Level-Based Feature | Description |
| --- | --- |
| Fundamental frequency ($f0$) | Acoustic equivalent to pitch; rate of vocal fold vibration; maximum of the autocorrelation function; models prosodic structure; speech melody indicator |
| Energy | Models intensity, on the basis of the amplitude in different intervals; average squared amplitude within a predefined time segment; stressing structure |
| Harmonics-to-noise ratio (HNR) | Spectral energy in voiced versus unvoiced segments; ratio between harmonic and aperiodic signal energy; breathiness indicator |
| Formant position ($F1$–$F6$) | Resonance frequencies of the vocal tract (VT) depending strongly on its actual shape; represent spectral maxima, and are known to model spoken content and speaker characteristics; influenced by lower jaw angle, tongue body angle, tongue body horizontal location, tongue tip angle, tongue tip horizontal location, relative lip height, lip protrusion, velum height |
| Formant bandwidth ($F\text{bw}1$–$F\text{bw}6$) | Models VT shape and energy loss of speech signal due to VT elasticity (yielding wall effect), viscoelasticity of VT tissue or heat-conduction-induced changes of air flow (jet streams, turbulences); width of the spectral band containing significant formant energy ($-3$-dB threshold) |
| Duration of voiced–unvoiced segments | Models temporal speech rhythm aspects such as speech rate and pause structure |
| Mel frequency cepstrum coefficients (MFCCs) | "Spectrum of the spectrum"; have been proven beneficial in speech emotion recognition and speech recognition tasks; homomorphic transform with equidistant band-pass filters on the mel scale; holistic and decorrelated representation of spectrum |
| Linear frequency cepstrum coefficients (LFCCs) | Similar to MFCCs but without the perceptual oriented transformation into the mel frequency scale; emphasize changes or periodicity in the spectrum, while being relatively robust against noise |
| Long-term average spectrum (LTAS) | Averages out formant information; giving general spectral trends; relative amount of energy within predefined frequency bands; speech quality |
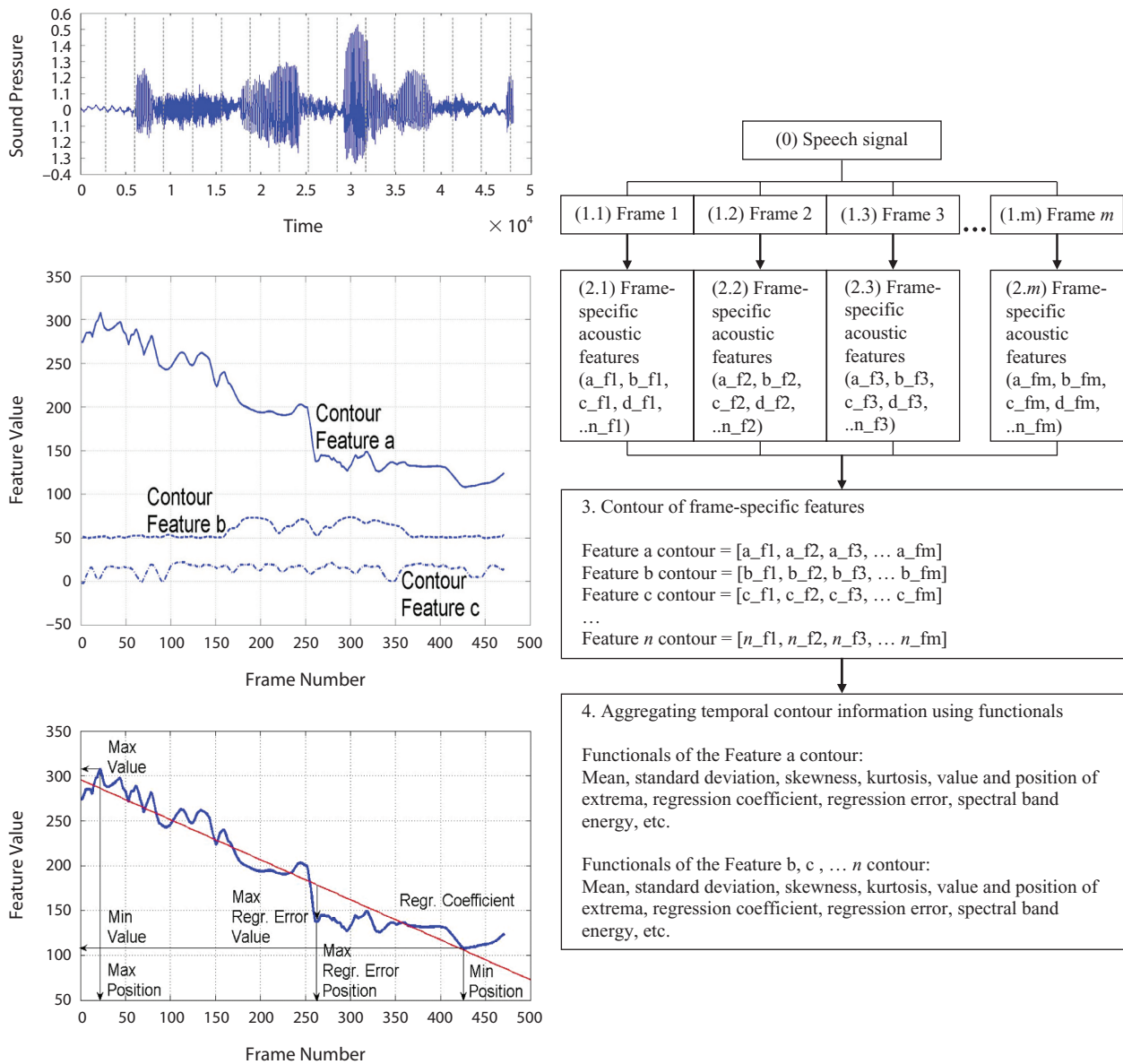
Figure 1. Processing flow of acoustic feature extraction: Step 1, framing; Step 2, frame-level descriptors (FLDs); Step 3, FLD contours; and Step 4, FLD contour functionals.

acoustic measurement process follows the speech-adapted steps of pattern recognition: (1) recording speech, (2) preprocessing, (3) feature computation, (4) dimensionality reduction, (5) classification, and (6) evaluation. The following listing gives a brief overview about possible variations in the measurement process.

**Recording speech**. *Source of verbal material*: size and domain of vocabulary; human-to-human, human-to-machine communication; monologue versus dialogue situations; speech databases (e.g., FAU Aibo Emotion Corpus—Batliner, Steidl, & Nöth, 2008; Sympafly—Batliner, Hacker, Steidl, Nöth, & Haas, 2003). *Speaking format*: vowel phonation, isolated words, connected speech, read speech, spontaneous speech. *Speech tasks*: referring to content, scheduling a business meeting, explain-

ing a route; referring to intensity and articulation-related speaking style, hyperarticulation, whispering, shouting. *Speaking style*: read speech, answering speech, command speech, descriptive speech, nonprompted speech, spontaneous speech. *Speech segment*: vowels, consonant types (fricative, stop, glide), consonant clusters, syllables, words, chunks, phrases. *Recording situation*: telephone recording, on-site recording, field recording, Wizard-of-Oz; noisy versus noise-subdued environment (e.g., driving with an open window vs. laboratory recording); rough versus clean speech signal quality (e.g., telephone call, radio communication vs. clean recording in 22.05 kHz, 16 bit).

**Preprocessing**. *Segmentation*: manual, (semi-)automatic segmentation (e.g., MAUS system; Schiel, 2004) of the speech signal in phonetic units of interest (e.g., specific vow-

els, types of consonants, or consonant clusters, stressed syllables, beginning or end of phrases). *Noise reduction*: outlier detection, moving average filter, low band-pass filter. *Framing and windowing*: size of frames (10–20 msec), degree of overlapping, window function (hamming, hanning).

**Feature computation**. *FLDs*: fundamental frequency, intensity, HNR, formant position, and bandwidth ($F1$–$F6$), linear predictive coding (LPC) coefficients, MFCCs, LFCCs, partitioning into voiced and unvoiced speech segments. *Functionals*: elementary statistics (e.g., linear moments, extrema values and positions, quartiles, ranges, length of time periods beyond threshold values, regression coefficients), spectral descriptors (e.g., spectral energy of low-frequency bands vs. high-frequency bands), and state space features (e.g., largest Lyapunov coefficient); automatic feature generation (genetic algorithms). *Normalization*: individual-speaker-specific baseline correction, age-/gender-specific normalization. *Acoustic information channels*: phonetic (prosody, articulation, speech quality), paralinguistic (affective bursts; e.g., deep breathing as a sound of relief). *Syntactic channels*: part-of-speech categories (e.g., number of nouns, noninflected adjectives), linguistic inquiry and word count (LIWC) word categories, and MRC psycholinguistic features (see Mairesse, Walker, Mehl, & Moore, 2007). *Semantic channels*: topic selection (e.g., self-focused, or pleasure talk; bag of words features), conversational behavior characteristics (listen vs. initiate conversation, back-channel behavior, formal style).

**Dimensionality reduction**. *Subset selection*: supervised filter-based (e.g., correlation, information gain ratio), unsupervised (e.g., density, entropy, salience), or wrapper-based subset selection (forward selection, backward elimination, sequential forward-floating search, genetic algorithm selection). *Feature transformation*: unsupervised (e.g., principal component analysis [PCA], PCA network, nonlinear autoassociative network, multidimensional scaling, kernel PCA, independent component analysis, Sammon's algorithm (nonlinear mapping), enhanced Lipschitz embedding, self-organizing map (SOM), spectral transformation, wavelet transformation); supervised (e.g., linear discriminant analysis).

**Classification**. *Classification granularity*: binary or multiple class prediction. *Classifier choice*: one-nearest neighbor, multilayer perceptron, support vector machine, linear discriminant analysis, hidden Markov model, decision tree, Gaussian mixture model. *Metaclassifier*: bagging, boosting, voting, stacking.

**Validation**. *Evaluation strategy*: $k$-fold cross-validation; leave-one-sample/speaker-out, multiple hold-out. *Reliability strategy*: recordings on different days for retest reliability (e.g., leave-one-session-out).

## METHOD

### Procedure, Subjects, and Speech Materials

We conducted a validation experiment to examine whether automatically trained models can be used to recognize the sleepiness of subjects. Our approach can be summarized in four steps: (1) Collect individual speech data and the associated sleepiness ratings for each subject; (2) extract relevant acoustic features from the speech data; (3) build statistical models of the sleepiness ratings based on the

acoustic features; (4) test the learned models on unseen speech data. The following sections describe each of these steps in more detail.

Twelve student volunteers took part in this study. Initial screening excluded those having severe sleep disorders or sleep difficulties. The subjects were instructed to maintain their normal sleep pattern and behavior. Due to recording and communication problems, the data of 2 subjects could be analyzed only in part (four speech samples excluded). We conducted a within-subjects sleep deprivation design (01.00–08.00 a.m.; seven 60-min sessions). Each session consisted of 40 min of driving in a real-car-driving simulation lab in nearly complete darkness, followed by responding to sleepiness questionnaires (2 min), recording speech material (2 min), conducting vigilance tests (15 min), and a break (1 min). During the night of sleep deprivation, a well-established, standardized subjective sleepiness questionnaire measure, the Karolinska Sleepiness Scale (KSS), was used by the subjects (self-assessment) and, additionally, by the two experimental assistants (outsider assessment). Time since sleep was at least 16 h, which was checked by wrist actometry. The subjects were prepared beforehand by simulator training. Several biosignals were recorded: EEG, EOG, ECG, and EMG. In addition, three video recordings (driver's head and pose, driver's eyes, driving scene) were stored. Also, several variables for the car, such as steering angle and lane deviation, were sampled, but their analysis will not be considered here. Further experimental details have been published elsewhere (Golz et al., 2007). In the version used in the present study, scores range from 1 to 10, and verbal descriptions are given for both even and odd scores: *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, can't stay awake* (10). Given the verbal descriptions, scores of 8 and higher appear to be most relevant from a practical perspective, since they describe a state in which the subject feels unable to stay awake. During the night, the subjects were confined to the laboratory, conducted a driving simulator task, 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, 16 bit). Furthermore, the subjects were given sufficient prior practice so that they were not uncomfortable with this procedure. The verbal material consisted of a simulated pilot–air-traffic-controller communication ("Cessna nine three four five Lima, county tower, runway two four in use, enter traffic pattern, report left base, wind calm, altimeter three zero point zero eight"). The subjects recorded other verbal material at the same session, but in this article, we will focus on the material described above. For training and classification purposes, the records were further divided into two classes: alert (A) and sleepy (S), with a threshold value of KSS $\geq 7.5$ (8 samples per subject; total number of speech samples, 94 samples; 34 A samples, 60 S samples; KSS = mean of the three KSS ratings; $M = 7.22$, $SD = 2.87$). The threshold has been validated by observations of microsleep events: Below threshold, we never observed any microsleep event.

### Feature Extraction

All acoustic measurements were taken utterance-wise, using the Praat speech analysis software for computing the FLDs (Boersma, 2001). As was mentioned above, we estimated the following 58 FLDs (see Figure 2): fundamental frequency, fundamental frequency peak process, intensity, harmonics-to-noise ratio, formant position and bandwidth ($F1$–$F6$), 15 LPCs, 12 MFCCs, 12 LFCCs, duration of voiced and unvoiced speech segments, and long-term average spectrum (LTAS). These 58 FLDs are joined by their first and second derivates (velocity and acceleration contours). Furthermore, these 174 speech feature contours are modeled in average by 129.56 functionals in time and frequency domain feature space.

1. Functionals from elementary statistics (*time domain*): min, max, range, mean, median, trimmed mean 10%, trimmed mean 25%, 10th, 25th, 75th, and 90th percentile, interquartile range, mean average deviation, standard deviation, skewness, kurtosis, robust regres-
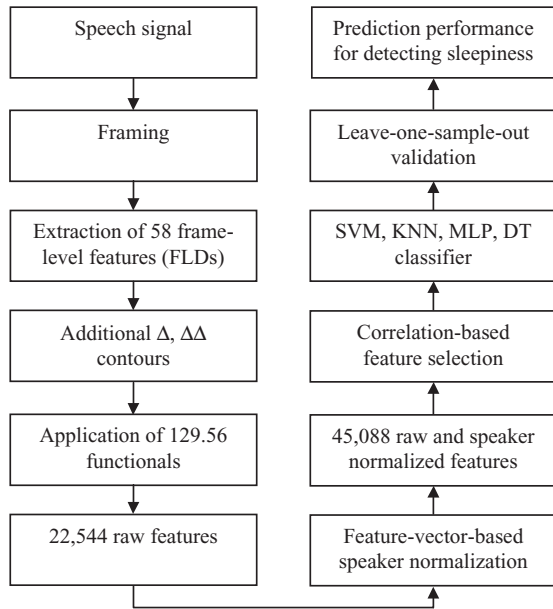
**Figure 2. Pattern-recognition-based processing flow of the complete measurement process, including the steps of recording, preprocessing, feature extraction, dimensionality reduction, and classification.**

sion coefficients, intercept, frequency of values beyond different thresholds (median $\pm$ 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 $*$ median), min and max position, relative min and max position; entropy, number of peaks, mean standard deviation, min and max of peak position, peak amplitude value, delta peak position, and delta peak amplitude.

2. Functionals from the *spectral domain*: spectral envelope (regression coefficient, intercept), power spectral density of five frequency bands, relative power, maximum within five frequency bands.

This procedure of combining FLDs and functionals results in 22,544 raw features. To take individual response patterns into account, we added the same number of speaker-normalized features (differences between raw feature vectors and the speaker-specific mean of this feature vector). In sum, we computed a total number of 45,088 features per speech sample.

### Dimensionality Reduction and Classification

The purpose of feature selection is to reduce the dimensionality that otherwise can hurt classification performance. The small amount of data also suggested that longer vectors would not be advantageous, due to overlearning of data. Optimization can be performed by (1) (un)supervised feature subset selection (e.g., correlation-filter-based or genetic-algorithm-based wrapper selection) and (2) (un)supervised feature transformation methods (e.g., principal component analysis, single-value decomposition, linear discriminant analysis). In this study, we used a relevance-maximizing rather than a redundancy-minimizing correlation filter approach (Pearson correlation > .40). This low-computational-effort-demanding technique leads to a compact representation of the feature space. Furthermore, wrapper-based supervised subset selections (the classifier's error serves as target function in a closed-loop procedure) were employed to optimize predicting accuracy, using search strategies as forward-selection-, backward-elimination-, and genetic-algorithm-based methods. These heuristic search algorithms are necessary because exhaustive search of the best feature subset is computationally prohibitive ($2^n - 1$ possible feature subset combinations; $n$ = number of features). In addition, we employed feature space transformation techniques for dimensionality reduction as principal component analysis, self-organizing map, and single-value decomposition.

**Support-vector machine (SVM) classifier**. After the rough feature selection, a classifier (modeling technique) can be designed using different approaches based on the concept of similarity (*k*-nearest neighbor [KNN]), probability (density estimators), or direct decision boundary construction (SVM). It still remains an intriguing challenge to develop algorithms that automatically select classifiers and parameters appropriate for a given data set. SVMs (Vapnik, 1995) have become popular among many different types of classification problems—for example, face identification, bioinformatics, and speaker recognition. The basic principle of this discriminative method is to find the best separating hyperplane between groups of datapoints that maximizes the margins. Referring to the classifier choice, one could consider the use of SVM here, since it has proven in many works to model static acoustic feature vectors best (Vlasenko et al., 2007). Thus, we used for the classification an SVM (dot kernel function). Due to the nonrelatedness of the SVM's generalization error to the input dimensionality, they offer a possibility to train generalizable, nonlinear classifiers in high-dimensional feature spaces, using a small training set; nevertheless, we applied several often used classifiers, as well as those that are described in the following.

**KNN classifier**. In order to classify an instance of test data into a class, the KNN classifier (Mitchell, 1997) calculates Euclidean distance metrics between test data and each instance of a training data set. Then the algorithm finds the $k$ closest training instances to the test instance. The class with the highest frequency among $k$ is the class mapped to the test data. These computationally demanding prototype-based classification rules are the key disadvantage of the KNN classifier when facing large sample sizes. Nevertheless, KNN classifers are always close to optimal in accuracy, for an appropriate choice of $k$. The best value for $k$ can only be determined empirically. We use a KNN with a typical number of neighbors ($k$ = 1, 2, or 3).

**Multilayer perceptron (MLP) classifier**. This popular type of feedforward neural network consists of an input layer (the input features), possibly several hidden layers of neurons, and an output layer. The neurons calculate the weighted sum of their input and compare it with a threshold to decide whether they should "fire." We chose an MLP of usual size and configuration (feedforward net, back-propagation, two hidden sigmoid layers, and five nodes within each layer). Moreover, several different classifiers were used in our study, such as a decision tree (DT), a random forest, a naive Bayes, a basic rule learner, a radial basis function (RBF), a logistic base, a fuzzy lattice reasoning, and a logistic regression. Due to data sparsity, a speaker-dependent approach was chosen: a leave-one-sample-out cross-validation; that is, in turn, one case was used as a test set and all others as train. The final classification errors were calculated by averaging over all classifications.

## RESULTS

### Relevance of Single Acoustic Features

The spectrogram in Figure 3 provides a first insight into possible sleepiness-sensitive acoustic features. As we can infer from the distances between the harmonics (white spectral maxima), the fundamental frequency of the alert speech sample (136 Hz) is higher than in the sleepy speech sample (129 Hz). This hint derived from the spectrogram can be proven over all speech samples ($r = -.42$). In analogy to the predictions of the cognitive-physiological mediator model proposed above, we found a decreased position of Formant 1 values for the sleepy speaker ($r = -.35$). In addition, the best three single features have the following significant correlations ($p < .05$) to sleepiness scores: *speaker-dependent mean-normalized frequency of values above 1,200 msec of the first derivate of the duration of voiced parts = .57; speaker-dependent mean-normalized average absolute deviation of intensity = .52; speaker-*
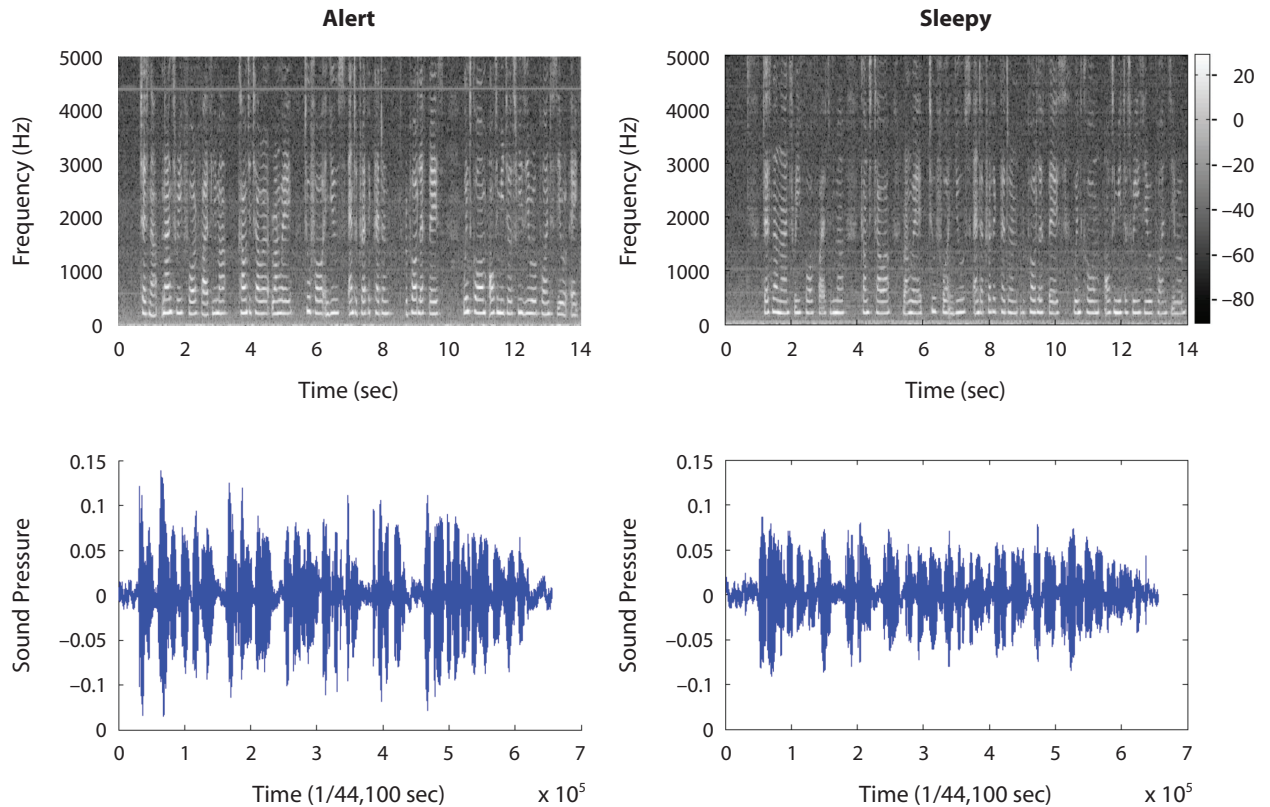
**Figure 3. Typical spectrograms and waveforms of the same utterance for an alert (left; Karolinska Sleepiness Scale [KSS] = 4) and a sleepy (right; KSS = 8) speaker. High-power spectral densities (PSDs) are coded white; low PSDs are colored gray.**

*dependent mean-normalized relative time beyond threshold of median + 2 * median of Formant 1 bandwidth = .51.*

### Dimensionality Reduction Methods

The removal of irrelevant and redundant features often improves the performance of classification algorithms. Following the standard pattern recognition procedure, we applied dimensionality reduction approaches as feature selection methods (filter and wrapper based) and unsupervised feature transformation methods. As is shown in Table 2, the most successful dimensionality reduction approach is the simple correlation-filter-based feature subset selection with 86.1% recognition rate and 82.8% class-wise averaged classification rate (RR = ratio of correctly classified samples divided by all samples, and CL = class-wise averaged classification rate).

### Overall Classification Results

In order to determine the multivariate prediction performance, different classifiers were applied on the 230 features remaining after the correlation filter procedure. For all configurations, we trained the classifiers and applied them on the unseen test sets (10-fold cross-validation). The averaged recognition rates (RR/CL) of the different classifiers for the two-class prediction problems are: SVM (86.1/82.8), MLP (80.9/79.3), 1-NN (73.4/70.3), 2-NN (62.8/69.5), 3-NN (76.6/72.1), DT (75.5/70.6), random forest (68.1/62.9), naive Bayes (73.4/70.9), basic rule learner (71.3/71.7), RBF (72.3/68.2), logistic base (86.1/82.4), fuzzy lattice reasoning (75.5/75.1), and logistic regression (86.2/82.4). The SVM prediction achieved the highest class-wise averaged classification rate ($r = .70$; ROC area under curve = .85), which reached significance, as compared with a pure chance-based classification [$\chi^2(1) = 45.5, p < .001$], and was therefore applied for further detailed FLD-based analyses.

### Feature Subset Comparison

It may be imaginable that not all feature subsets are equally important for the sleepiness measurement. The features remaining after the correlation filter procedure can help to answer this question. The number of features (FLD classes) surviving this selection procedure are as follows: fundamental frequency = 3; duration voiced/unvoiced = 39; MFCCs = 19; LFCCs = 72; LPCs = 67; HNR/intensity = 20; formants = 8, LTAS = 2. Referring to the feature set size surviving the filter procedure, the LFCCs are the most successful feature subset. An alternative approach to determining the relative importance of feature subsets is to apply classifiers on each feature subset separately. The results depicted in Table 3 show the relevance of the formants, $f0$, and LFCC feature subset.

### DISCUSSION

To cover possible prosodic, speech quality, and articulatory changes in sleepy speech, an uncommonly large 45k

**Table 2**
**Recognition Rates (RRs) and Class-Wise Averaged Classification Rates**
**(CLs) for Different Dimensionality Reduction Techniques Using the**
**Support-Vector Machine (SVM) As Classifier and**
**Leave-One-Sample-Out Validation Strategy**

| Feature Transformation | # | Feature Selection | # | RR | CL |
|---|---|---|---|---|---|
| PCA (90% variance) | 36 | – | – | 77.6 | 73.6 |
| PCA (99% variance) | 61 | – | – | 76.6 | 72.7 |
| SVD | 2 | – | – | 72.3 | 66.2 |
| SVD | 3 | – | – | 79.8 | 74.6 |
| SVD | 5 | – | – | 77.6 | 72.3 |
| SOM | 2 | – | – | 63.8 | 50.0 |
| – | – | Forward selection SVM | 14 | 74.5 | 71.1 |
| – | – | Backward elimination SVM | 124 | 76.6 | 70.2 |
| – | – | Genetic algorithm SVM | 37 | 83.0 | 77.7 |
| PCA (90% variance) | 36 | Genetic algorithm SVM | 21 | 78.7 | 71.8 |
| SVD | 3 | Genetic algorithm SVM | 2 | 72.3 | 64.3 |
| – | – | Correlation filter | 230 | 86.1 | 82.8 |

Note—#, number of remaining components.

feature space was reduced with several dimensionality reduction methods and then fed into diverse classifiers. The main findings of the present study may be summarized as follows. First, acoustic features extracted from read speech and modeled with pattern recognition methods contain a substantial amount of information about the speaker's sleepiness state. Second, the uncommonly large brute force feature set (45k) computation was able to determine whether a subject's sleepiness was beyond a critical threshold. In our experiments on this two-class classification problem, we achieved an accuracy rate of over 86% on unseen data but known speakers (70% on unknown speakers) with an SVM classifier.

The most important FLD feature classes for this prediction were the following: (1) according to the sum of features remaining after the correlation filter, LFCCs, LPCs, and duration of voiced/unvoiced; (2) according to the prediction accuracy of the single FLD feature class, formants, $f0$, and LFCCs. These results are mainly in accord with the predictions of the cognitive-physiological mediator model

**Table 3**
**Recognition Rates (RRs) and Class-Wise Averaged**
**Classification Rates (CLs) (in Percentages) on the Test Set, Using**
**Different Frame-Level Descriptor (FLD) Feature Sets (Raw and**
**Speaker-Normalized Features Surviving the Correlation Filter)**
**on the Support-Vector Machine Classifier**

| FLDs | # | Raw | | # | Raw and Normalized | |
|---|---|---|---|---|---|---|
| | | RR | CL | | RR | CL |
| Formants | 2 | 71.3 | 65.4 | 8 | 86.2 | 82.8 |
| $f0$ | 2 | 72.3 | 68.1 | 3 | 78.7 | 75.7 |
| LFCCs | 18 | 73.4 | 70.3 | 72 | 77.7 | 72.9 |
| MFCCs | 5 | 72.3 | 67.5 | 19 | 74.5 | 69.2 |
| LPCs | 14 | 74.5 | 71.1 | 67 | 70.2 | 65.8 |
| HNR/Intensity | 11 | 70.2 | 65.8 | 20 | 66.0 | 60.0 |
| Duration | 1 | 64.9 | 57.8 | 39 | 64.9 | 56.0 |
| LTAS | 0 | – | – | 2 | 67.0 | 54.4 |
| All FLDs | 53 | 70.2 | 65.8 | 230 | 86.1 | 82.8 |

Note—#, number of features; LFCCs, linear frequency cepstrum coefficients; MFCCs, mel frequency cepstrum coefficients; LPCs, linear predictive coding coefficients; HNR, harmonics-to-noise ratio; LTAS, long-term average spectrum.

of sleepiness (e.g., the *relaxed open mouth* phenomenen referred to above in a speaker-normalized setting), which highlights the importance of fundamental frequency, formants, and duration features. Especially the eight formant–FLD features are unexpectedly successful in the classification task, reaching a classification accuracy comparable to that for the full-size feature set (230 features). Using this full-size feature set containing all FLDs, we achieved on this two-class classification problem a recognition rate of over 86% on unseen but speaker-dependent data with an SVM classifier. Our classification performance is in the same range as that obtained for comparable tasks—for example, for emotional user state classification (see Batliner et al., 2006; Schuller et al., 2008; Vlasenko et al., 2007).

**Limitations and Future Work**

The validity of our results is limited by several facts. First, the major criticism refers to the choice of the applied ground truth. The used fusion of one self-report and two observer report measures could be criticized because of its (semi-)subjective nature, lacking an involvement of "objective" physiological ground truth measures. Until now, many studies have shown associations between physiological data (e.g., EEG or EOG) and fatigue. Nevertheless, they still remain in a premature developmental stage, without offering a standardized, commonly accepted scaling, as realized in the KSS. Furthermore, the KSS has proven its validity in several studies, particularly when the application context has been unlikely to provoke self- or observer deception. Despite this currently premature status of *behavioral and physiological sleepiness instruments*, they offer a promising potential as future "gold standard" measures. Second, sleepiness might be confounded by annoyance states due to the multiple repetition of the speaking task. Thus, the results obtained in the present study with a within-subjects design should be replicated with a between-subjects design. Third, the present results are preliminary and need to be replicated using a natural speech environment: It would seem advisable that future studies address the main topics of improving

the acoustic sleepiness analysis and finding evidence for its validity in real-world applications.

With the limiting factors described above in mind, our present findings should be viewed as preliminary ones that warrant more controlled research. A short sketch (see Figure 4) will sum up possible starting points for future research facing primarily the challenge of improving the measurement precision of the acoustic sleepiness detection.

1. *Recording*. Collecting sleepy speech samples from different types of speakers and speaking styles would provide a broadly diversified learning data set to compute submodels for different sleepiness responder groups and different confounder–sleepiness constellations (e.g., a submodel for a sleepy speaker with a head cold). Thus, collecting different sleepy speech samples within an open-source speech database would provide an infrastructural research background that would enhance further progress in acoustic sleepiness analysis. Emotion and stress speech databases (e.g., FAU Aibo Emotion Corpus; Batliner et al., 2008) could serve as a model for this kind of open-source speech corpora.

2. *Preprocessing*. Finding and segmenting sleepiness-sensitive phonetic units (phons, consonant clusters, or syllables in different word and phrasal unit positions) could improve the overall detection rates, especially in free and spontaneous speech with unrestricted vocabulary size and domain. Automatic speech recognition could serve for this purpose. In addition, a word tagger could help to count semantic and syntactic categories that might provide information about sleepiness as well (e.g., part-of-speech classes; Batliner et al., 1999).

3. *Feature extraction*. Signal-processing features derived from state space domains—such as average angle or length of embedded space vectors and recurrence quantification analyses—should be computed, and feature transformation should be applied (Webber & Zbilut, 1994). Moreover, evolutionary feature generation methods could be used to find further features. 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. In addition, hierarchical functionals (Schuller

et al., 2008) might help to identify sleepiness-sensitive subparts within a speech segment.

4. *Dimensionality reduction*. For finding the optimal feature subset, further supervised filter-based subset selection methods (e.g., information gain ratio) or supervised wrapper-based subset selection methods (e.g., sequential forward floating search, genetic algorithm selection) should be applied. Other methods for reducing the dimensionality of the feature space are unsupervised feature transformations methods (e.g., PCA network, nonlinear autoassociative network, multidimensional scaling, independent component analysis, Sammon map, enhanced Lipschitz embedding, SOM) or supervised feature transformation methods (e.g., LDA).

5. *Classification*. A third class should be added to the classification task, serving as a warning stage within a sleepiness detection system. Furthermore, future work on sleepiness prediction could consider metaclassifier methods such as bagging, boosting, or stacking (Wolpert, 1992), including exhaustive parameter optimizations. Dividing between male and female classification models might be as promising, such as applying maximum-likelihood Bayes classifiers or fuzzy membership indexing.

## REFERENCES

Batliner, A., Hacker, C., Steidl, S., Nöth, E., & Haas, J. (2003). User states, user strategies, and system performance: How to match the one with the other. In *Proceedings of an ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems* (Vol. 1, pp. 5-10).

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R., & Niemann, H. (1999). Automatic annotation and classification of phrase accents in spontaneous speech. In *Proceedings of the European Conference on Speech Communication and Technology* (Vol. 6, pp. 519-522).

Batliner, A., Steidl, S., & Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo Emotion Corpus. In *Proceedings of a Satellite Workshop of*
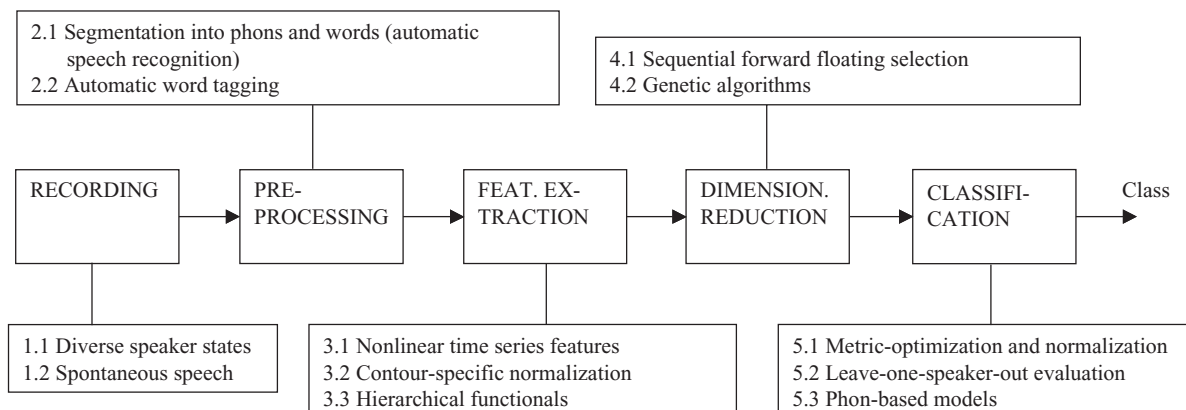


**Figure 4. Possible developments within the field of acoustic sleepiness detection.**

*LREC 2008 on Corpora for Research on Emotion and Affect* (Vol. 1, pp. 28-31).

BATLINER, A., STEIDL, S., SCHULLER, B., SEPPI, D., LASKOWSKI, K., VOGT, T., ET AL. (2006). Combining efforts for improving automatic classification of emotional user states. In T. Erjavec & J. Z. Gros (Eds.), *Language technologies, IS-LTC 2006* (pp. 240-245). Ljubljana, Slovenia: Infornacijska Druzba.

BOERSMA, P. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, **5**, 341-345.

BRATZKE, D., ROLKE, B., ULRICH, R., & PETERS, M. (2007). Central slowing during the night. *Journal of Psychological Science*, **18**, 456-461.

DAVIDSON, P. R., JONES, R. D., & PEIRIS, M. T. (2007). EEG-based behavioral microsleep detection with high temporal resolution. *IEEE Transactions on Biomedical Engineering*, **54**, 832-839.

DINGES, D. F., & KRIBBS, N. (1991). Performing while sleepy: Effects of experimentally induced sleepiness. In T. H. Monk (Ed.), *Sleep, sleepiness and performance* (pp. 97-128). Chichester, U.K.: Wiley.

DURMER, J. S., & DINGES, D. F. (2005). Neurocognitive consequences of sleep deprivation. *Seminars in Neurology*, **25**, 117-129.

ENGLE-FRIEDMAN, M., RIELA, S., GOLAN, R., VENTUNEAC, A. M., DAVIS, C. M., JEFFERSON, A. D., & MAJOR, D. (2003). The effect of sleep loss on next day effort. *Journal of Sleep Research*, **12**, 113-124.

FLATLEY, D., REYNER, L. A., & HORNE, J. A. (2004). *Sleep-related crashes on sections of different road types in the UK (1995–2001)* (Road Safety Research Rep. No. 52, pp. 4-132). London: Department of Transport.

GOLZ, M., SOMMER, D., HOLZBRECHER, M., & SCHNUPP, T. (2007). Detection and prediction of driver's microsleep events. In RS4C (Eds.), *Proceedings of the 14th International Conference of Road Safety on Four Continents*. Bangkok, Thailand.

GREELEY, H. P., BERG, J., FRIETS, E., WILSON, J., GREENOUGH, G., PICONE, J., ET AL. (2007). Fatigue estimation using voice analysis. *Behavior Research Methods*, **39**, 610-619.

HARRISON, Y., & HORNE, J. A. (1997). Sleep deprivation affects speech. *Sleep*, **20**, 871-877.

HARWOOD, K., BARNETT, B., & WICKENS, C. D. (1988). Situational awareness: A conceptual and methodological framework. In F. E. McIntire (Ed.), *Proceedings of the 11th Biennial Psychology in the Department of Defense Symposium* (pp. 7-23). Colorado Springs: U.S. Air Force Academy.

HORBERRY, T., HUTCHINS, R., & TONG, R. (2008). *Motorcycle rider fatigue: A review*. (Road Safety Research Rep. No. 78, pp. 4-63). London: Department for Transport.

HORNE, J. A. (1988). Sleep loss and "divergent" thinking ability. *Sleep*, **11**, 528-536.

INGRE, M., ÅKERSTEDT, T., PETERS, B., ANUND, A., & KECKLUND, G. (2006). Subjective sleepiness, simulated driving performance and blink duration: Examining individual differences. *Journal of Sleep Research*, **15**, 47-53.

JENNINGS, J. R., MONK, T. H., & VAN DER MOLEN, M. W. (2003). Sleep deprivation influences some but not all processes of supervisory attention. *Psychological Science*, **14**, 473-479.

KIENAST, M., & SENDLMEIER, W. F. (2000). Acoustical analysis of spectral and temporal changes in emotional speech. In *Proceedings of the ISCA Workshop on Speech and Emotion* (pp. 92-97). Bonn: ISCA.

KRAJEWSKI, J. (2008). *Acoustic sleepiness analysis.* Unpublished doctoral thesis, University of Wuppertal.

LEVELT, W. J. M., ROELOFS, A., & MEYER, A. S. (1999). A theory of lexical access in speech production. *Journal of Behavioral & Brain Sciences*, **22**, 1-75.

LINDE, L., & BERGSTRÖM, M. (1992). The effect of one night without sleep on problem-solving and immediate recall. *Psychological Research*, **54**, 127-136.

MAIRESSE, F., WALKER, M., MEHL, M., & MOORE, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, **30**, 457-500.

MELAMED, S., & OKSENBERG, A. (2002). Excessive daytime sleepiness and risk of occupational injuries in non-shift daytime workers. *Sleep*, **25**, 315-322.

MITCHELL, T. M. (1997). *Machine learning.* New York: McGraw-Hill.

NILSSON, J. P., SODERSTROM, M., KARLSSON, A. U., LEKANDER, M., ÅKERSTEDT, T., LINDROTH, N. E., & AXELSSON, J. (2005). Less effective executive functioning after one night's sleep deprivation. *Journal of Sleep Research*, **14**, 1-6.

NWE, T. L., LI, H., & DONG, M. (2006). Analysis and detection of speech under sleep deprivation. In *Proceedings of Interspeech* (Vol. 9, pp. 17-21). Bonn: ISCA.

O'SHAUGHNESSY, D. (2000). *Speech communications: Human and machine.* Piscataway, NJ: IEEE Press.

READ, L. (2006). Road safety Part 1: Alcohol, drugs and fatigue. In Department for Transport (Ed.), *Road safety Part 1* (pp. 1-12). London: Department for Transport.

ROGERS, N. L., DORRIAN, J., & DINGES, D. F. (2003). Sleep, waking and neurobehavioral performance. In J. M. Kreuger (Ed.), *Frontiers in Bioscience 8* (pp. 1056-1067). Albertson, NY: Frontiers in Bioscience.

SCHIEL, F. (2004). MAUS goes iterative. In *Proceedings of the IV. International Conference on Language Resources and Evaluation* (pp. 1015-1018).

SCHLEICHER, R., GALLEY, N., BRIEST, S., & GALLEY, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired? *Ergonomics*, **51**, 982-1010.

SCHULLER, B., BATLINER, A., SEPPI, D., STEIDL, S., VOGT, T., WAGNER, J., ET AL. (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Proceedings of Interspeech* (Vol. 8, pp. 2253-2256). Bonn: ISCA.

SCHULLER, B., WIMMER, M., MÖSENLECHNER, L., KERN, C., & RIGOLL, G. (2008). Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 33, pp. 4501-4504).

SOMMER, D., CHEN, M., GOLZ, M., TRUTSCHEL, U., & MANDIC, D. (2005). Fusion of state space and frequency domain features for improved microsleep detection. In W. Dutch et al. (Eds.), *Proceedings of the International Conference on Artifical Neural Networks (ICANN 2005)* (pp. 753-759). Berlin: Springer.

TARTTER, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, **27**, 24-27.

TASSI, P., PELLERIN, N., MOESSINGER, M., ESCHENLAUER, R., & MUZET, A. (2000). Variation of visual detection over the 24-hour period in humans. *Journal of Chronobiology International*, **17**, 795-805.

VAPNIK, V. N. (1995). *The nature of statistical learning theory.* New York: Springer.

VLASENKO, B., SCHULLER, B., WENDEMUTH, A., & RIGOLL, G. (2007). Combining frame and turnlevel information for robust recognition of emotions within speech. In *Proceedings of Interspeech* (Vol. 8, pp. 2249-2252). Bonn: ISCA.

VÖHRINGER-KUHNT, T., BAUMGARTEN, T., KARRER, K., & BRIEST, S. (2004). Wierwille's method of driver drowsiness evaluation revisited. In *International Conference on Traffic and Transport Psychology* (Vol. 3, pp. 5-9).

WEBBER, C. L., & ZBILUT, J. P. (1994). Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, **76**, 965-973.

WESENSTEN, N. J., BELENKY, G., THORNE, D. R., KAUTZ, M. A., & BALKIN, T. J. (2004). Modafinil vs. caffeine: Effects on fatigue during sleep deprivation. *Aviation, Space, & Environmental Medicine*, **75**, 520-525.

WHITMORE, J., & FISHER, S. (1996). Speech during sustained operations. *Speech Communication*, **20**, 55-70.

WILHELM, B., GIEDKE H., LÜDTKE, H., BITTNER, E., HOFMANN, A., & WILHELM, H. (2001). Daytime variations in central nervous system activation measured by a pupillographic sleepiness test. *Journal of Sleep Research*, **10**, 1-7.

WOLPERT, D. H. (1992). Stacked generalization. *Journal of Neural Networks*, **5**, 241-259.

WRIGHT, N., & MCGOWN, A. (2001). Vigilance on the civil flight deck: Incidence of sleepiness and sleep during long-haul flights and associated changes in physiological parameters. *Ergonomics*, **44**, 82-106.

ZILS, E., SPRENGER, A., HEIDE, W., BORN, J., & GAIS, S. (2005). Differential effects of sleep deprivation on saccadic eye movements. *Sleep*, **28**, 1109-1115.