Medium-term speaker states—A review on intoxication, sleepiness and the first challenge

Björn Schuller[a,b,*], Stefan Steidl[c,d], Anton Batliner[a,d], Florian Schiel[e], Jarek Krajewski[f], Felix Weninger[a], Florian Eyben[a]

[a] *Technische Universität München, Institute for Human–Machine Communication, Germany*
[b] *Joanneum Research Forschungsgesellschaft mbH, DIGITAL – Institute for Information and Communication Technologies, Austria*
[c] *ICSI, Berkeley, CA, USA*
[d] *FAU Erlangen-Nuremberg, Pattern Recognition Lab, Germany*
[e] *Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universität München, Germany*
[f] *University of Würzburg, Industrial and Organizational Psychology, Germany*

Abstract

In the emerging field of computational paralinguistics, most research efforts are devoted to either short-term speaker states such as emotions, or long-term traits such as personality, gender, or age. To bridge this gap on the time axis, and hence broaden the scope of the field, the INTERSPEECH 2011 Speaker State Challenge addressed the algorithmic analysis of medium-term speaker states: alcohol intoxication and sleepiness, both of which are highly relevant in high risk environments. Preserving the paradigms of the two previous INTERSPEECH Challenges, researchers were invited to participate in a large-scale evaluation providing unified testing conditions. This article reviews previous efforts to automatically recognise intoxication and sleepiness from speech signals, and gives an overview on the Challenge conditions and data sets, the methods used by the participants, and their results. By fusing participants' systems, we show that binary classification of alcoholisation and sleepiness from short-term observations, i.e., single utterances, can both reach over 72% accuracy on unseen test data; furthermore, we demonstrate that these medium-term states can be recognised more robustly by fusing short-term classifiers along the time axis, reaching up to 91% accuracy for intoxication and 75% for sleepiness.

Keywords: Computational paralinguistics; Intoxication; Sleepiness; Survey; Challenge

1. Medium-term speaker states—an introduction

Amongst paralinguistic phenomena, personality — characterising human beings — and emotions — characterising cognitive, psycho-physiological experiences of human beings — are arguably most prototypical. A quick web

* Corresponding author at: Joanneum Research Forschungsgesellschaft mbH, DIGITAL – Institute for Information and Communication Technologies, Austria. Tel.: +43 316 876 5012; fax: +49 89 289 28535.
E-mail addresses: bjoern.schuller@joanneum.at, schuller@tum.de, schuller@IEEE.org (B. Schuller).

search[1] yielded these frequencies for the following noun phrases: "personality traits": 8,760,000, "emotional states": 3,600,000, "personality states": 81,100, and "emotional traits": 121,000. On the time axis, personality can be described as constituted by long-term traits, and emotions as short-term states; our crude web search corroborates this statement because traits most often go together with personality, and emotions with states. However, there are enough personality states and emotional traits to make one believe that there is something in between: Personality forms the pre-condition for specific emotions, as stated by Revelle and Scherer (2009) who speak of: "[. . .] the notion of habitual or trait emotionality; that is, an individual difference variable consisting of a disposition to experience certain types of emotions more frequently than other people." In the same way, personality can manifest itself in specific states. (Note that for the sake of the argument, we disregard the sloppy use of 'state' as a sort of synonym for 'trait'.) Now the two phenomena we deal with in this article — intoxication and sleepiness — certainly have to do both with personality and with emotion: Certain personality traits may predispose an individual to drug use, cf. Kalivas (2003), Loukas et al. (2000) and Echeburúa et al. (2007), and under the influence of drugs, the manifestation of emotion can vary, and the personality can change as well in the long run. A similar relationship can be observed for sleepiness disorders, cf. Sforza et al. (2002). Both states can be described and processed on their own, without reference to personality traits that favour or disfavour them. Moreover, intoxication and sleepiness are definitely medium-term, not short-term states, lasting normally at least several minutes in the case of sleepiness and several hours in the case of intoxication—definitely not only a few seconds: It takes some time to get drunk or sleepy, and it gets some time to get sober or awake again. Moreover, in some pathological cases, we can speak of long-term, permanent intoxication, and of habitual sleepiness that should be taken care of.

Medium-term states can be self-induced, such as intoxication—apart from specific circumstances (for instance, knock-out drops that are consumed unintentionally). Or, they can be partly self-induced, such as sleepiness—sometimes the situation requires to stay awake. To mention other medium-term states: In the conceptualisation of Scherer (2003), affective states such as mood, interpersonal stances, and attitudes are medium-term as well. Such affective states are normally not self-induced but a result of complex interactions of dispositions with the surroundings, especially in communicative situations. Intuitively, health states are mostly medium-term as well, and are partly self-induced, partly not. In our context, all these states are of course mostly relevant if we can diagnose them with the help of speech parameters.

So far, the bulk of research within computational paralinguistics has been devoted to the two 'endpoints' on the time scale, i.e., to long-term traits and to short-term states. Within automatic speech recognition, mostly traits such as age and gender were addressed, arguably because pertinent information is straightforward and easy to get, and it can be employed within envisaged applications. If it comes to the modelling and processing of human–human or human–machine interaction, both personality traits and especially affective/emotional states have been frequently dealt with. The focus of research efforts has not been on the phenomena in between, i.e., on medium-term states. One of the reasons is arguably the difficulty to collect such data within tightly controlled experimental settings: It is very easy to tell people to act in some emotional way, and it is relatively easy to elicit emotions in an experimental setting. Every human being has a personality, thus the problem is not primarily to find, elicit, and segment it. In contrast, the experimental effort is much higher and ethical considerations play a much greater role if it comes to medium-term states such as sleepiness and especially intoxication. Given the fact that alcoholic intoxication and sleepiness are pivotal factors in accidents, it would be highly desirable to model and detect them automatically, based on nonintrusive recordings of speech. To this aim, the INTERSPEECH 2011 Speaker State Challenge addressed these two medium-term states within the same strictly controlled paradigm as has been used for the preceding two INTERSPEECH challenges which addressed emotional states and several types of traits.

From a processing point of view, medium-term states can and have to be handled differently from short term states and long term traits: The segmentation of the speech signal into coherent chunks, cf. Batliner et al. (2010), is not critical because the state does not change fast as in the case of emotions. This means in turn, however, that we cannot investigate different medium term states within one straightforward experimental setting. Recordings have to be made over periods of several hours or days when we want to monitor changes. However, in contrast to personality traits, we can monitor individuals before, during, and after intoxication or sleepiness, in medium-term time frames as well. On

[1] Google search on February 12, 2012.

the one hand, all this makes experimentation more cumbersome; on the other hand, we will see that this opens new possibilities to collect cumulative evidence.

The remainder of this article is organised as follows: In Sections 2 and 3, we review previous studies relating speech signal parameters to intoxication and sleepiness levels, respectively. We then describe the Challenge framework and results in Section 4. Our conclusions are drawn in Section 5.

2. Speaker intoxication—a review

2.1. Earlier studies

It is a widely accepted hypothesis that alcoholic intoxication (AI), as other factors such as fatigue, stress and illness, influences the way a person speaks. Quite a number of studies during the last decades have investigated this hypothesis from different points of view: looking for reliable acoustic (Künzel and Braun, 2003; Cooney et al., 1998) or behaviouristic (Hollien et al., 2001; Behne et al., 1991; Sobell et al., 1982; Trojan and Kryspin-Exner, 1968) features that may indicate intoxication, studying the physiological effects of alcohol on the articulators (Watanabe et al., 1994) or even pursuing forensic questions (Künzel and Braun, 2003; Braun, 1991; Klingholz et al., 1988; Martin and Yuchtman, 1986) such as in the infamous case of the captain of the Exxon Valdez: When the oil tanker Exxon Valdez stranded in Alaska in 1989, the captain of the ship was suspected of being under alcoholic influence during the time of the crisis. Forensic analysis of the recorded air traffic indicated that the spectra of the phone /s/ were skewed in the direction of an /S/ sound[2] which was considered as an indicator for drunkenness (Johnson et al., 1990).

2.2. Automatic detection of AI

Alcoholic intoxication has always been and still is one of the major causes for traffic accidents. AI can forensically be measured by (ordered by descending reliability): taking blood samples (blood alcohol concentration, BAC), breath alcohol detectors (breath alcohol concentration, BRAC) and a variety of tests often applied in experimental and cognitive psychology (mainly reaction time and motor control). All these tests can only be applied either in random checks or post-accidentally, that is after an accident has already happened. Currently there are no known practical methods to routinely check on the AI of a driver pre-emptively.

On the other hand, nowadays automobiles are equipped with a growing number of functions controlled by speech input. Prominent examples are entertainment (radio, CD), control of the hands-free telephone, and input to the navigation system. The type of speech applied here is typical command and control consisting of a limited number of pre-determined commands (often only 3–7 words) and issued to the car system after pressing a button on the steering wheel via a built-in microphone in the roof of the cabin. However, it is to be expected in the near future that more sophisticated voice input in the form of keyword activation and free (continuous) speech — as already demonstrated in prototype systems — will be incorporated into standard car systems. This leads to the interesting question whether a pre-emptive test of alcoholic intoxication using speech input might be feasible in the automotive environment: Since the driver of an automobile will increasingly use his or her voice to communicate with the car system, it could theoretically be possible for the car system to automatically retrieve indicators for AI and react accordingly, for instance by warning the driver about her or his condition.

The detection of AI (and other speaker states) with automatic methods differs from classic pattern recognition tasks where the training or enrollment data matches the test data and the subject is sober when producing both. In the AI detection application, it is not possible to collect speech data from the intoxicated speaker during the enrollment. The subject is usually sober when producing the enrollment data and either sober or intoxicated in the test situation, which makes the recognition task more difficult (Schiel et al., 2010). To our knowledge nobody has yet approached such a realistic detection problem (involving out-of-the-lab language and a variety of speakers of different age and gender). Aside from reports within the Interspeech Speaker State Challenge 2011 based on the ALC (Schuller et al., 2011), only Levit et al. (2001) and Sigmund et al. (2010) have reported about attempts to detect the grade of intoxication from the speech signal by means of statistical classification.

[2] Phonetic symbols in SAM-PA.

In Levit et al. (2001) prosodic features were used in binary classification by an Artificial Neural Network. Read speech was recorded from 33 German male speakers and labelled with 'below 0.8 per mill' and 'above 0.8 per mill' BAC. Due to data sparsity the tests were performed on the validation set. The best combination of recognisers yielded about 69% recognition rate on the binary classification.

Sigmund et al. (2010) analysed the glottal pulses of the vowels /a/, /e/, /i/, /u/ and /o/ of young Czech male speakers using Iterative Adaptive Inverse Filtering. In a linear classification scheme into two classes 'sober' and 'intoxicated' (BAC >1.0 per mill), the glottal pulse features derived from the vowels /e/ and /o/ yielded the best results (76.6% and 77.0% respectively) on a group of 12 male test speakers.

2.3. Realistic situation

For a realistic evaluation of automatic AI detection, data outside the laboratory environment have to be investigated. Unfortunately, most of the studies listed above have the following in common:

- they dealt with read speech in an acoustically clean environment only,
- they analysed speech of male speakers only,
- they analysed less than 40 speakers,
- the AI was not measured reliably, and finally
- the analysed empirical speech data have not been made available for other research groups so that different methods of detection could be compared.

The speech data used in the Interspeech 2011 Speaker State Challenge has been selected to remedy some of these deficiencies (see Section 4.2 for details).

2.4. Human performance

In this context the question arises what performance is to be expected as a gold standard in the task of AI detection. One straightforward way in scientific studies is to use the measured BAC which ideally should be provided for each investigated speaker. On the other hand it is possible that some speakers are able to mask their AI perfectly and therefore produce speech signals indistinguishable from normal (sober) speech. Hence, it would be interesting to see how human listeners perform on the same task, since for many recognition tasks concerning speech, humans are considered to perform better than machines.

Based on common experience, most listeners claim that they can reliably recognise AI in the speech uttered by intoxicated persons. In a number of earlier studies, results of identification tests on laboratory speech of intoxicated speakers have been reported. In Martin and Yuchtman (1986), 44 male subjects performed a forced choice test on 192 sentences read by 8 (male) speakers resulting in an identification rate of 62.5%. In another study Klingholz et al. (1988) reported a recognition rate of 54.0% on 30 s of read text spoken by 11 male speakers intoxicated with <1.0 per mill BAC and judged by 12 listeners; recognition rates increased to a maximum of 82.0% when the BAC was >1.0 per mill. In Künzel et al. (1992) 33 male speakers produced read and semi-spontaneous speech under varying intoxication levels. 10–12 s long stimuli derived from these recording were used in an identification task performed by 30 listeners yielding an average recognition rate of 66.8%; recognition rates increased linearly from 50.0 to 96.0% with increasing BAC (estimated from BRAC) over a range of 0.4–2.0 per mill.

If the performance of human listeners is significantly above chance, which features do they use for their (successful) decisions and is there a difference between female and male listeners? Another question relates to the speaker dependency on the AI detection task. More specifically: are there distinctive speaker groups that

- reveal their AI more easily,
- mask their AI better than others, or
- appear to be under AI although being sober?

In analogy to Doddington's 'zoo of speaker verification' (Doddington, 1998) we could label these three groups as *lambs, wolves* and *goats*.

A small class-balanced and length-normalised sample drawn from 16 speakers of the corpus of the Intoxication Sub-Challenge (cf. Section 4.2.1) was used in a simple forced choice perception experiment to quantify the ability of human listeners to distinguish between sober and intoxicated speech. The average discrimination rate (unweighted accuracy) of 47 listeners was significantly above chance with 71.65%, but still far from the optimum. The listeners were more successful in detecting intoxication in female voices than in male voices, and in read rather than in spontaneous speech. On the other hand, female and male listeners showed the same detection performance. Prosodic information could be exploited by human listeners for the decision process but probably not as much as other types of features. There was some evidence that AI detection is strongly influenced by the individual behaviour of speakers. More specifically, some speakers were easier recognised than the average (lambs), while some speakers were even judged to be intoxicated when in fact being sober (goats). No significant indication for speakers that can mask their intoxication perfectly (wolves) were found in this study (Schiel, 2011).

In Ultes et al. (2011) a large balanced sample of 3600 recordings from the Intoxication Sub-Challenge was subdivided into chunks of 50 recordings each and presented to 79 listeners in an AI identification task. The listeners had to make a hard decision between 'sober' and 'intoxicated' for each presented recording. They achieved 55.8% unweighted accuracy in this task which is only slightly above chance. No effects of listener gender, of listener age (2 age groups below/above 50 years) nor of the length of the speech sample on the detection accuracy were found in this study (speaker gender was not tested). Inter-rater kappa was very low with $\kappa = 0.15$ indicating that the raters had great difficulties with this task. Although it is to be expected that human raters perform better in a discrimination task than in identification, it is still surprising that the performance in this study is so much lower than in Schiel (2011).

3. Speaker sleepiness—a review

Regarding speaker sleepiness — the second medium-term speaker state addressed in the Challenge besides intoxication — we now discuss potential applications of automatic analysis, acoustic correlates of sleepiness, and algorithmic approaches.

3.1. Automatic detection of sleepiness

Sleepiness is a crucial factor in a variety of incidents and accidents in road traffic (Flatley et al., 2004; Horberry et al., 2008; Read, 2006) and work contexts (e.g., safety sensitive fields such as chemical factories, nuclear power stations, and air traffic control: Melamed and Oksenberg, 2002; Wright and McGown, 2001). For instance, 21% of the reported incidents mentioned in the Aviation Safety Reporting System (including pilots and air traffic controllers) were related to sleepiness. Thus, the prediction and warning of traffic employees against impending critical sleepiness play 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 detection of and automatic countermeasures to sleepiness. Knowing the speaker's sleepiness state can contribute to the naturalness and acceptance of HCI. If the user shows unusual sleepiness, 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 sleepiness-impaired attentional and cognitive resources.

Hence, many efforts have been reported in the literature for measuring sleepiness related states (Fulda and Popp, 2011; Golz et al., 2005; Sommer et al., 2009; Schnupp et al., 2009; Krajewski et al., 2010). These approaches have focused mainly on measures of

- the autonomous nervous system such as pupil size (Schnieder et al., 2012), eye blinking (Schleicher et al., 2008), or heart rate (Heinze et al., 2009),
- the central nervous system such as EEG (Sommer et al., 2009), and
- behavioural expression data such as steering behaviour, tracking tasks, gross body movement (Krajewski et al., 2010; Schenka et al., 2010; Schnupp et al., 2009)

in order to characterise the sleepiness state.

But these electrode-based (EOG/EEG reaching 15% error rate; Golz et al., 2005) or video-based instruments (PERCLOS reaching 32% error rate; Sommer et al., 2009) still do not meet the demands of an everyday life measurement system (Golz et al., 2010). The major drawbacks are

- 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
- a lack of comfort and longevity due to electrode sensor application.

In contrast to these electrode- or video-based instruments, the utilisation 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 and 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 (Schuller et al., 2010).

3.2. Acoustic correlates of sleepiness

The temporary states of sleepiness show a distinct pattern of effects on speech, despite various influencing factors from the acoustic environment to interdependencies with other states and traits (cf. Table 2 in Traunmüller, 2000). Sleepiness related cognitive-physiological changes can influence voice characteristics indirectly, according to the following stages of speech production (O'Shaughnessy, 2000): At the stage of cognitive speech planning, a reduced cognitive processing speed might lead to impaired speech planning (Levelt et al., 1999) and impaired neuromuscular motor coordination processes, slowing down the transduction of neuromuscular commands into articulator movement and affecting the feedback of articulator positions (Bratzke et al., 2007; Dinges and Kribbs, 1991). At the stage of muscular actions, 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 (Ananthapadmanabha, 2011). 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 (Story, 2002).

Thus, it can be anticipated that sleepy speech compared with rested speech might exhibit acoustic changes in

- prosody—such as monotonic and flattened intonation, shifted speech rate, or reduced syllable duration due to, e.g., slowed cognitive speech planning,
- articulation—such as slurred, less crisp pronunciation, mispronunciations, abrupt articulatory changes, speech errors, or hesitations due to, e.g., impaired motor coordination processes and aversion of spending compensatory effort (Lieberman et al., 1995), and
- speech quality—such as tensed, nasal, or breathy speech due to, e.g., impaired coordination of velum closure (Kostyk and Rochet, 1998).

Furthermore, linguistic changes in syntactic and semantic structures can be expected (e.g., simplified grammatical structure; favouring easy accessible, frequently used words; repetitive, vague, and rambled speech; favouring words associated with the field of deactivation or negative mood; less involved and polite discourse behaviour, less backchannels). These changes — summarised in the cognitive-physiological mediator model of sleepiness induced speech changes (Krajewski and Kroeger, 2007; Krajewski et al., 2012) — 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. Previous studies have analysed mostly highly artificial speech material (meaningless syllable list; Vollrath, 1993), only small sample sizes ($N < 20$), and small phonetic feature sets (Harrison and Horne, 1997; Shahidi et al., 2010; Vogel et al., 2010; Whitmore and Fisher, 1996) or small feature sets containing only perceptual acoustic features (e.g., pitch, intensity, speech rate). Signal processing based features, well-known from speech and speaker recognition (e.g., mel-frequency cepstrum coefficients, MFCCs) have received little attention

(Greeley et al., 2007; Krajewski and Kroeger, 2007; Krajewski et al., 2009; Nwe et al., 2006; Zhang et al., 2010). Moreover, often no sleepiness scaled reference instruments are applied (e.g., psychophysiological measures or psychomotorical tests, Dhupati et al., 2010), and only very long time-since-sleep periods (> 24 h) were analysed, which narrows the range of potential application scenarios.

In detail, the following sleepiness induced changes of speech parameters have been reported:

- a decreased mean fundamental frequency (F0) (Johannes et al., 2000; Krajewski et al., 2009; Nwe et al., 2006) vs. an increased mean fundamental frequency (Ruiz et al., 2010),
- a decreased standard deviation of F0 (Morris et al., 1960; Nwe et al., 2006) vs. an increased standard deviation of F0 (Vogel et al., 2010),
- a decreased speech rate (Morris et al., 1960; Vogel et al., 2010), respectively an increased mean pause length (Vogel et al., 2010),
- a decreased ratio of duration of voiced-unvoiced parts (derived from Mel-Frequency Cepstrum Coefficients and Gaussian Mixture Model; Dhupati et al., 2010),
- an increase in misarticulations (Morris et al., 1960) vs. no change in misarticulations (Harrison and Horne, 1997),
- increased speaking errors (Morris et al., 1960) vs. no changes in increased speaking errors (Harrison and Horne, 1997),
- a decreased Formant 1 position (Krajewski et al., 2009),
- a decreased Formant 4 Variation (Vogel et al., 2010),
- an increased time of high values for Formant 1 bandwidth (Krajewski et al., 2009),
- an increased average absolute deviation of intensity (Krajewski et al., 2009),
- a decrease of the slope of the long term average spectrum (Krajewski et al., 2012), and
- a decreased fractal dimension, maximum of the Cao's minimum embedding dimensions (Krajewski et al., 2012).

These partly ambiguous results could be explained by various methodological issues such as small sample sizes, different choices of speech tasks (read, automated, vowel phonation vs. spontaneous speech; e.g., contradictory effects on speech rate for read vs. automated speech), or different distribution of sleepiness intensity. Similarly to the issues mentioned for alcoholised speech, the speech data used in the Interspeech 2011 Speaker State Challenge has been selected to overcome some of these drawbacks.

3.3. Automatic speech sleepiness detection

In order to show a comprehensive summary of the results achieved so far in automatic speech sleepiness detection, Table 1 is presented. To ensure generalisability, the sample size was evaluated as well as the naturalness of the speech, as indicated by the speech task mentioned. Speaker independent modelling is listed due to different demands of application scenarios. The recording protocol is given to estimate the sleepiness distribution and ease of the classification task. Finally, we provide unweighted accuracy to evaluate the overall performance of the applied feature sets and classifier.

The studies mainly recorded only a few speakers ($N < 25$), and only several seconds of speech material per speaker. Moreover, the naturalness of the speech scenario measured by the applied speech task (mainly vowel phonation or command and control speech) can be considered as restricted. A further restriction of the listed studies is given by their speaker dependent modelling, which does not meet the demands of several relevant application scenarios. As can be derived from the given recording protocols, several studies choose a very high time-since-sleep, inducing very strong sleepiness. On the one hand this distribution of sleepiness simplifies the classification task, on the other hand, again the range of application scenarios is narrowed. The size of the applied feature sets is in the range 0.1–45 k, often containing F0–F5, MFCCs or LPCCs. Comparing the applied classifiers (e.g., RF, MLP, LDA, or HMM), no consistent results could be achieved favouring one classifier. The presented unweighted average recall rates show speaker-dependent rates of about 80–85%. In sum, small sample sizes, irrelevant high time-since-sleep values, speaker-dependent modelling, and non-comparable sleepiness reference values narrowed down the generalisability of the results found so far.

Table 1
Comparison of sleepiness studies involving automatic speech detection. (1), Greeley et al. (2007); (2), Krajewski et al. (2009); (3), Krajewski and Kroeger (2007); (4), Krajewski et al. (2012); (5), Nwe et al. (2006); (6), Zhang et al. (2010); #S, number of speakers; Lang., language; en, English; ge, German; ma, Mandarin; read, read speech; c+c, command and control speech; spon, spontaneous monolog/dialog; SI, speaker independence; TSS, time-since-sleep [h] of recording protocol (minimum:stepsize:maximum); MFCC, mel frequency cepstrum coefficients; LFCC, linear frequency cepstrum coefficients; LPCC, linear predictive coding coefficients; PHSC, Pitch and Harmonic frequency Spectral Coefficients; F0, fundamental frequency; F0–F5, formant 0–5 frequency position and bandwidth; MLP, multi layer perceptron, PNN, probabilistic neural network; LDA, linear discriminant analysis; HMM, hidden markov model; GMM, gaussian mixture model; LR, logistic regression; kNN, *k*-nearest neighbor; DT, decision tree; RF, random forest; BN, Bayesian network; UA, unweighted accuracy.

	#*S*	Lang.	Speech task	SI	TSS	Features	Classifier	UA [%]
(1)	2	en, ge	read (31 words)	?	4:4:34	MFCC (36)	GMM	?
(2)	12	ge	c+c (1 sentence)	no	12:1:20	F0–F5, MFCC, LFCC, LPCC (45,088)	KNN, MLP, SVM	82.8
(3)	23	ge	c+c (1 sentence)	no	12:1:20	F0–F5, HNR, MFCC (169)	MLP, LDA	84.2
(4)	77	en	vowel phonation	no	12:1:20	nonlin. dynamics (395), 169 features from (3)	MLP, LDA, LR, kNN, DT, RF, BN	79.6
(5)	12	en	spon (DCIEM, Bard et al., 1996)	?	0.5/60	PHSC, MFCC, LPCC	HMM	86.5
(6)	1	ma	vowel phonation (6 vowels)	no	0:6:18	LPCC, MFCC	PNN	?

4. The first challenge on medium-term speaker states: intoxication and sleepiness

Paralinguistics comprises much more than, on the one hand, emotional states which can change in a short time, and on the other hand, speaker-specific traits such as gender or age that normally either do not change at all or only over a longer period of time. Thus, the INTERSPEECH 2011 Speaker State Challenge broadened the scope by addressing two less researched speaker states, by that focusing on the crucial application domain of security and safety: the computational analysis of intoxication and sleepiness in speech. Apart from intelligent and socially competent future agents and robots, main applications are found in the medical domain and surveillance in high-risk environments such as driving, steering or controlling (Brenner and Cash, 1991).

The INTERSPEECH 2011 Speaker State Challenge, organised by Björn Schuller (TUM, Germany), Stefan Steidl (ICSI, USA), Anton Batliner (FAU, Germany), Florian Schiel (University of Munich, Germany), and Jarek Krajewski (University of Wuppertal, Germany) was held in conjunction with INTERSPEECH 2011 in Florence, Italy, 28–31 August 2011. This Challenge was the first open public evaluation of speech-based speaker state recognition systems aimed at medium-term speaker states — namely intoxication and sleepiness — in between short term states such as emotion or interest and long-term traits such as age or gender. As in previous Challenges organised by the first three organisers starting with INTERSPEECH 2009, strict comparability was given: The German Alcohol Language Corpus (provided by the Bavarian Archive for Speech Signals).[3] Please contact bas@bas.uni-muenchen.de or refer directly to the BAS catalogue at www.bas.uni-muenchen.de/Bas. and the Sleepy Language Corpus — both containing real affection of the speaker by either alcohol intoxication or sleep deprivation — served as a basis with clearly defined test, training, and development partitions incorporating speaker independence as needed in most real-life settings. The first consists of 39 h of speech, stemming from 154 speakers in gender balance, and serves to evaluate features and algorithms for the estimation of speaker intoxication in gradual blood alcohol concentration (BAC). The second features 21 h of speech recordings of 99 subjects, annotated in the 10 different levels of sleepiness defined by the Karolinska Sleepiness Scale (KSS). The verbal material is of different complexity reaching from sustained vowel phonation to natural communication.

Two Sub-challenges were addressed in the 2011 Speaker State Challenge, each using two classes. In the *Intoxication Sub-Challenge*, the alcoholisation of a speaker had to be determined as two-class classification task: *alcoholised* for a

[3] The ALC is available for unrestricted scientific and commercial usage. Interested parties may obtain copies of the full corpus at BAS (BAS distribution fees apply.)

Table 2
Baseline feature set provided for the Challenge, based on low-level descriptors extracted on frame level and functionals applied on recording level.

(a) 60 provided low-level descriptors (LLD)
4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
50 spectral LLD
RASTA-filt. auditory spectrum, bands 1–26 (0–8 kHz)
MFCC 1–12
Spectral energy 25–650 Hz, 1 k–4 kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Flux, Entropy, Var., Skewness, Kurtosis, Slope
5 voice related LLD
F0, Probability of voicing
Jitter (local, delta)
Shimmer (local)
(b) 33/6 applied functionals.
33 base functionals
Quartiles 1–3
3 inter-quartile ranges
1% percentile (≈min), 99% percentile (≈max)
Percentile range 1–99%
Arithmetic mean, standard deviation
Skewness, kurtosis
Mean of peak distances
Standard deviation of peak distances
Mean value of peaks
Mean value of peaks — arithmetic mean
Linear regression slope and quadratic error
Quadratic regression a and b and quadratic error
Contour centroid
Duration signal is below 25% range
Duration signal is above 90% range
Duration signal is rising/falling
Gain of linear prediction (LP)
LP coefficients 1–5
6 F0 functionals
Percentage of non-zero frames
Mean, max, min, std. dev. of segment length
Input duration in seconds

BAC exceeding 0.5 per mill[4] or *non-alcoholised* for a BAC equal or below 0.5 per mill. In the Sleepiness Sub-Challenge, sleepiness of speakers had to be determined for a level exceeding level 7.5 on the KSS reaching from one (extremely alert) to 10 (cannot stay awake). This threshold (between level 7, 'sleepy, some effort to stay awake' and level 8, 'very sleepy, great effort to stay awake') has been validated by observations of microsleep events: Below this threshold, we have never observed any microsleep events; further, from level 7 to level 8 there is a significant increase in the accident risk (Ingre et al., 2006). Before presenting the intoxication and sleepiness tasks in detail, let us first outline the set of acoustic features that was provided to the Challenge participants and was used to compute the baseline classification accuracies for the Challenge (Table 4).

[4] Per mill BAC by volume, which is standard in most central and eastern European countries; further ways exist, e.g., percent BAC by volume, i.e., the range resembles 0.028–0.175 per cent (Australia, Canada, USA), points by volume (GB), per mill by BAC per mass (Scandinavia) or part per million.

Table 3
Challenge partitions of ALC. 'NAL' denotes recordings of non-alcoholised, i.e., BAC per mill in the interval [0;0.5], and 'AL' recordings of alcoholised speakers, i.e., BAC per mill in]0.5;1.75].

#ALC	NAL	AL	Total
Train	3750	1650	5400
Develop	2790	1170	3960
Test	1620	1380	3000
Train + Develop	6540	2820	9360
Train + Develop + Test	8160	4200	12,360

Table 4
Challenge partitions of SLC. 'NSL' denotes recordings of non-sleepy, i.e., KSS in the interval [1;7.5], and 'SL' recordings of sleepy speakers, i.e., KSS]7.5;10].

#SLC	NSL	SL	Total
Train	2125	1241	3366
Develop	1836	1079	2915
Test	1957	851	2808
Train + Develop	3961	2320	6281
Train + Develop + Test	5918	3171	9089

4.1. Challenge features

In this Challenge, an extended set of features with respect to the INTERSPEECH 2009 Emotion Challenge (384 features) (Schuller et al., 2009) and INTERSPEECH 2010 Paralinguistic Challenge (1 582 features) (Schuller et al., 2010a) is given to the participants, again using the open-source Emotion and Affect Recognition (openEAR) (Eyben et al., 2009) toolkit's feature extracting backend openSMILE (Eyben et al., 2010). The feature set consists of 4368 features comprising features known as relevant for these tasks (Chin and Pisoni, 1997; Dhupati et al., 2010) built from three sets of low-level descriptors (LLDs) extracted on frame level and one corresponding set of functionals for each LLD set, applied on recording level, i.e., to entire audio files.

The LLD sets are given in Table 2a: A major novelty concerning LLD compared to the previous Challenge set (Schuller et al., 2010a) is the auditory spectrum derived loudness measure and the use of RASTA-style filtered auditory spectra instead of Mel-spectra, as well as a slightly extended set of statistical spectral descriptors (such as entropy, variance, etc.). These features were added to have their potential evaluated in the Challenge setting. The new loudness measure serves as a better measure of perceptual loudness than the linear or logarithmic signal energy. Loudness variations are an important descriptor for speech prosody, thus it is desirable to have a measure for this which fits human perception well. The RASTA-style filtered auditory spectra are introduced for two reasons: (a) an auditory weighting is applied to the mel-band spectra to better model the ear's frequency perception, and (b) the RASTA-style filtering reduces the influence of stationary as well as highly instationary background and non-speech sounds on the spectra, as the time domain filter emphasises frequencies in the 4–8 Hz region, the syllable rate of speech. The statistical spectral descriptors describe the shape of the spectral energy distribution over the frequency axis. They would have distinctly different values for voiced and unvoiced spectra for example. Yet, as they describe the general shape of the spectrum, they do not contain the same information as the probability of voicing, for example, which is computed by the pitch tracker based on the strength of the fundamental frequency and its harmonics.

Further, a base set of 33 functionals is introduced as shown in Table 2b. Again, compared to the previous set, the use of autoregressive model coefficients as functionals is new. We use the autocorrelation method to compute linear predictive coding (LPC) coefficients and gain from low level descriptor contours. While in speech coding the purpose of LPC is to identify the vocal tract transfer function and the vocalic formant structure, when applying LPC as a functional to arbitrary data contours, there is generally no interpretation regarding the human voice production. Yet, the coefficients of the AR model represent a measure of how correlated adjacent samples are. They allow us to discriminate if the signal behaves predictable or not in a local region. Moreover, the standard deviation of the intra-peak distances was added as a functional. Assuming that peaks in a data contour correspond to certain important points, such

as prosodic emphasis if they are loudness maxima or pitch maxima, for example, they carry important information for emotion related phenomena. The same holds for their distribution over the time axis, whose regularity is roughly described by the standard deviation of the distances between the peaks.

In the set of functionals applied to the spectral and energy related LLDs, the standard deviation of the segment lengths is new as well. Also, a new algorithm for splitting the contour into segments is used. Previously this was based on delta thresholding, where a new segment was started when the signal rose by a pre-defined relative (to the signal's range) amount in a short time frame. Now, a new segment boundary is given each time the LLD's value (after simple moving average filtering with 3 frames width) crosses the values (min + 0.25·range) and (min + 0.75·range). This gives a more stable and meaningful segmentation, which does not rely on large jumps, but identifies regions where the signal remains continuously within certain bounds. As with the standard deviation of the inter peak distances, the standard deviation of the segment lengths describes the temporal regularity of the signal.

To the 54 energy and spectral LLDs and their first order deltas, the base functional set and the mean, max, min, and the standard deviation of the segment length are applied, resulting in 3996 features. To the five pitch and voice quality LLDs and their first order deltas, the base functional set as well as the quadratic mean and the rise and fall durations of the signal are applied only to voiced regions (probability of voicing greater 0.7). This adds another 360 features. Another 12 features are obtained by applying a small set of six functionals to the F_0 contour (including non-voiced regions where F0 is set to 0) and its first order derivative as also shown in Table 2b. Please note that segments in this case correspond to continuous voiced regions, i.e., where F_0 is >0. The configuration for the extraction of the features with openSMILE is also provided and allows, e.g., to use the LLD on frame basis, or to alter and add features.

4.2. Intoxication Sub-Challenge

4.2.1. Alcohol Language Corpus (ALC)

A brief description of the ALC project is given in this section. For a detailed description of the corpus please refer to Schiel and Heinrich (2009) and Schiel et al. (2012).

ALC comprises 162 speakers (84 male, 78 female) within the age range 21–75, mean age 31.0 years and standard deviation 9.5 years, from 5 different locations in Germany. To obtain a gender balanced set, 154 speakers (77 male, 77 female) are selected randomly for the Challenge; these are further randomly partitioned into gender balanced training, development and test sets according to Table 3.

Speakers voluntarily underwent a systematic intoxication test supervised by the staff of the Institute of Legal Medicine, Munich. Before the test, each speaker chose the BAC he/she wanted to reach during the intoxication test. Using both Watson- and Widmark formula (Schiel et al., 2012), the amount of required alcohol for each person was estimated and handed to the subject. After consumption, the speaker waited another 20 min before undergoing a BRAC test and a blood sample test (BAC). For the Challenge, only the BAC value is considered. The possible range is between 0.28 and 1.75 per mill. Immediately after the tests, the speaker was asked to perform the ALC speech test which lasted no longer than 15 min, to avoid significant changes caused by fatigue or saturation/decomposition of the measured blood alcohol level. At least two weeks later the speaker was required to undergo a second recording in sober condition, which took about 30 min. Both tests took place in the same acoustic environment and were supervised by the same member of the BAS staff, who also acted as the conversational partner for dialogue recordings. The speech signal was recorded with two different microphones: a headset Beyerdynamic Opus 54.16/3 and an AKG Q400 mouse microphone, frequently used for in-car voice input, located in the middle of the front ceiling of the automobile. For the Challenge, only the headset microphone is considered; signals are down-sampled from 44.1 kHz to 16 kHz sampling rate. Further, for the Challenge only the following meta data associated with each recording are provided: speaker ID, gender, and BAC (not for test). All speakers are prompted with the same material. Orthographic (in extended SpeechDat format) and phonologic transcripts as well as an automatic phonetic segmentation of all spoken items are provided for all recordings in Praat's TextGrid and BAS Partitur Format (BPF). Three different speech styles are part of each ALC recording session: read speech, spontaneous speech, and command & control. Speech styles are not marked for the Challenge.

4.2.2. Baseline results

For transparency and easy reproducibility, we use the WEKA data mining tool kit for classification (Hall et al., 2009), as we did for the INTERSPEECH 2009 Emotion Challenge and the 2010 Paralinguistic Challenge. As classifier

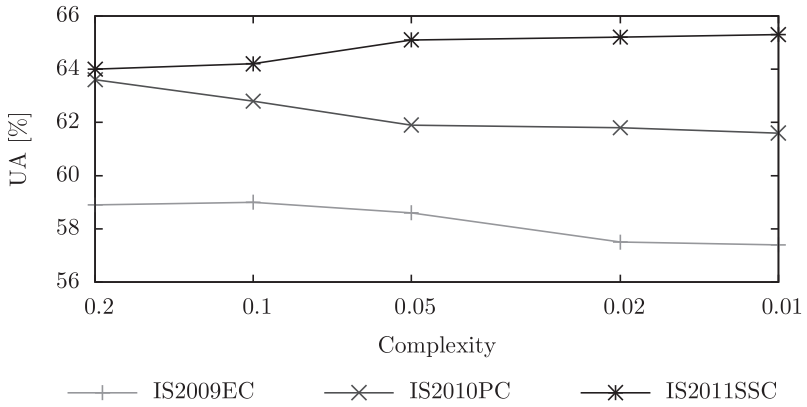


Fig. 1. Optimisation of SVM complexity by unweighted accuracy (UA) on the development partition of the ALC when training on the training partition after SMOTE. Three different feature sets are evaluated (cf. Table 7).

we chose Support Vector Machines (SVM) with linear kernel functions and Sequential Minimal Optimisation (SMO) as learning algorithm. We performed a limited grid search on the development set to find a suitable value for the complexity, which influences the number of support vectors for the hyperplane construction. We further use WEKA's implementation of the Synthetic Minority Over-sampling Technique (SMOTE), introduced by Chawla et al. (2002), as was done for the INTERSPEECH 2009 Emotion Challenge baseline, to balance instances in the learning partitions. If training and development partitions are united, SMOTE is applied subsequently to the unification. The results of the SVM complexity optimisation when training on the training partition of ALC and testing on the development partition is shown in Fig. 1 in terms of UA—the Challenge competition measure. We further evaluate the former feature sets of the 2009 and 2010 Challenges in comparison to the one provided for this Challenge. As can be seen, the new feature set prevails throughout the considered complexity range, especially in comparison to the 2009 Emotion Challenge feature set which mainly consists of basic MFCC and pitch functionals. Based on the optimal complexity $C = .01$ as found on the development partition, Table 5 shows baseline results for the *Intoxication Sub-Challenge* by UA and WA. As the distribution among classes is not balanced, the competition measure is UA as earlier stated. Results are given for training on the train partition and testing on the development partition—this could be freely done by participants, as well as for training on the unification of the training and development partitions and testing on the test partition—these results could be uploaded five times by the participants.

4.3. Sleepiness Sub-Challenge

4.3.1. Sleepy language corpus

99 participants took part in six partial sleep deprivation studies. The mean age of subjects was 24.9 years, with a standard deviation of 4.2 years and a range of 20–52 years. The recordings took place in a realistic car environment or in lecture-rooms (sampling rate 44.1 kHz, down-sampled to 16 kHz, quantisation 16 bit, microphone-to-mouth distance 0.3 m). The speech data consisted of different tasks: isolated vowels, i.e., sustained vowel phonation, sustained loud

Table 5
Intoxication Sub-Challenge baseline results by unweighted and weighted accuracy (UA/WA). SMO learned pairwise SVM with linear kernel, complexity optimised on development partition to 0.01. SMOTE on (united) learning instances. Feature sets IS 2009 EC, IS 2010 PC, and IS 2011 SSC correspond to the official sets of the Challenges (Emotion, Schuller et al., 2009, Paralinguistic, Schuller et al., 2010a, and Speaker State, Schuller et al., 2011) held at INTERSPEECH in these years.

[%]	Train vs. Develop		Train + Develop vs. Test	
Features	UA	WA	UA	WA
IS 2009 EC	57.4	65.3	60.3	60.2
IS 2010 PC	61.6	66.1	63.2	62.6
IS 2011 SSC	65.3	69.2	**65.9**	66.4

The final baseline is set in bold face.

vowel phonation, and sustained smiling vowel phonation; read speech: “Die Sonne und der Nordwind” (the story of ‘the North Wind and the Sun’, widely used within phonetics, speech pathology, and alike); commands/requests: 10 simulated driver assistance system commands/requests in German, e.g., “Ich suche die Friesenstraße” (‘I am looking for the Friesen street’); four simulated pilot-air traffic controller communication statements; moreover, a description of a picture and giving a PowerPoint guided, but non-scripted 20 minutes presentation in front of 50 listeners. A well established, standardised subjective sleepiness questionnaire measure, the Karolinska Sleepiness Scale, was used by the subjects (self-assessment) and additionally by the two experimental assistants (observer assessment, given by assessors who had been formally trained to apply a standardised set of judging criteria). In the version used in the present study, scores range from 1 to 10: extremely alert (1), very alert (2), alert (3), rather alert (4), neither alert nor sleepy (5), some signs of sleepiness (6), sleepy, but no effort to stay awake (7), sleepy, some effort to stay awake (8), very sleepy, great effort to stay awake, struggling against sleep (9), extremely sleepy, cannot stay awake (10). Given these verbal descriptions, scores greater than 7.5 appear to be most relevant from a practical perspective as they describe a state in which the subject feels unable to stay awake. For training and classification purposes, the recordings (mean = 5.9, standard deviation = 2.2) were thus divided into two classes: not sleepy (‘NSL’) and sleepy (‘SL’) samples with the threshold of 7.5 (ca. 94 samples per subject; in total 9277 samples). Besides these categorical labels, the SLC features speaker meta data (i. e., speaker ID and gender), and multiple annotation tracks of sleepiness ratings. Neither orthographic nor phonologic transcripts are provided for the recordings.

The six partial sleep deprivation studies can be described as follows:

- Study 1 has a within-subject, partial sleep deprivation design (8 pm–4 am) located in a noise subdued lab, hourly recordings (ca. 95 samples per subject; in total 2570 samples), 10 male and 15 female subjects, maximum time-since-sleep (tss) = 20 h, KSS = 5.18 $\pm$ 2.02.
- Study 2 has a within-subject, partial sleep deprivation design (8 pm–8 am) located in a driving simulator, nearly hourly recordings (ca. 118 samples per subject; in total 1411 samples), 9 male and 3 female subjects, max. tss = 24 h, KSS = 7.52 $\pm$ 2.28.
- Study 3 has a within-subject, partial sleep deprivation design (8 pm–2 am, 5 days) located in private home settings, hourly recordings (ca. 280 samples per subject; in total 3361 samples), 5 female subjects, max. tss = 18 h, KSS = 6.24 $\pm$ 2.37.
- Study 4 has a within-subject design (10 am–6 pm) located in a noise subdued lab, two recording sessions with a time lag of 8 h (40 samples per subject; in total 720 samples), 7 male and 11 female subjects, max. tss = 10 h, KSS = 4.31 $\pm$ 1.66.
- Study 5 has a within-subject design (2–5 pm) located in a realistic car environment, one recording sessions before and one after a three hour drive (40 samples per subject; in total 920 samples), 12 male and 11 female subjects, tss = 12 h, KSS = 5.13 $\pm$ 2.29. Study 6 has a single recorded presentation session (10–12 am) located in lecture-rooms, hourly recordings (ca. 15 samples per subject; in total 302 samples), 5 male and 11 female subjects, max. tss = 5 h, KSS = 3.20 $\pm$ 0.68. Further details of the studies can be found in Krajewski et al. (2009, 2012) and Krajewski and Kroeger (2007).

To follow a straightforward protocol for partitioning the SLC into the Challenge sets, the available turns were first divided into males (m) and females (f) per study. Then, the turns from male and from female subjects were split speaker-independently, in ascending order of subject ID, into training, development, and test instances. This subdivision not only ensures speaker-independent partitions, but also provides for stratification by gender and study setup (environment and degree of sleep deprivation). Out of the 99 subjects, 36 (20 f, 16 m) were assigned to the training, 30 (17 f, 13 m) to the development, and 33 (19 f, 14 m) to the test set. For the purpose of the Challenge, all turns including linguistic cues on the sleepiness level (e.g., “Ich bin sehr müde” — “I’m very tired”) were removed from the test set — 188 in total. The distribution of instances is given in Table 4.

In Table 6, we show the inter-rater agreement between the self (S) and observer assessments (O_i, $i = 1, 2, 3$) of the subjects’ sleepiness levels. We provide unweighted (Cohen’s) κ as well as κ^1 (weighted by the absolute disagreement on the KSS scale) and κ^2 (weighted by the squared disagreement). In Table 6a, Cohen’s κ is computed for a hypothetical nominal assessment (sleepy or non-sleepy) derived from the ordinal ratings in accordance with the derivation of the Challenge instance labels (‘non-sleepy’: KSS 1–7; ‘sleepy’: KSS 8–10); this ‘conversion’ is done since we believe that Cohen’s κ for the ten-point KSS scale would underestimate the degree of agreement. Overall, we observe sufficiently

Table 6
Inter-rater agreement on SLC, and agreement with consensus (i.e., binary Challenge label for κ and rounded mean of O_i for κ^1, κ^2). S: self assessment. O_i, O_j: observer asssessments.

Ratings	κ	κ^1	κ^2
(a) Inter-rater agreement			
S $\leftrightarrow$ O_1	.693	.738	.908
S $\leftrightarrow$ O_2	.655	.669	.872
S $\leftrightarrow$ O_3	.745	.685	.885
Mean (S $\leftrightarrow$ O_i)	.698	.697	.888
O_1 $\leftrightarrow$ O_2	.586	.659	.875
O_1 $\leftrightarrow$ O_3	.662	.691	.886
O_2 $\leftrightarrow$ O_3	.740	.699	.883
Mean (O_i $\leftrightarrow$ O_j)	.663	.683	.881
Mean	.680	.690	.885
(b) Agreement with rater consensus			
S	.829	.832	.955
O_1	.755	.814	.950
O_2	.801	.787	.940
O_3	.870	.805	.948
Mean	.814	.809	.948

high κ, κ^1 and κ^2 values for inter-rater agreement ($\kappa \geq .586$, $\kappa^1 \geq .659$; $\kappa^2 \geq .872$). The fact that the weighted κ variants indicate higher agreement than unweighted κ can be attributed to ratings being generally close to each other on the KSS scale. Furthermore, we observe that the average agreement of the self-assessment with the observer assessments ($\kappa = .698$) is slightly higher than the average agreement of observer assessments with each other ($\kappa = .663$); this justifies weighting the self-assessment and observer assessments equally in calculating the Challenge label.

In Table 6b, we proceed to quantify the agreement of individual assessments with the 'consensus' in terms of κ statistics. More precisely, we compute Cohen's κ of the nominal individual ratings with the Challenge label, which is obtained by discretising the mean ordinal rating as indicated above ('non-sleepy': mean KSS ≤ 7.5; 'sleepy' otherwise); the κ^1 and κ^2 measurements are taken between the ordinal KSS ratings and the mean KSS rating, rounded to the closest integer number in $\{1, \ldots, 10\}$. Generally, we obtain κ values widely above.7 regardless of the κ variant considered; this particularly indicates high agreement of the raters with the Challenge label (mean $\kappa = .814$). Finally, we note that mean κ^2 is observed as high as.948.

4.3.2. Baseline

The baseline for the Sleepiness Sub-Challenge was computed in full analogy to the Intoxication Sub-Challenge: An optimal combination of feature set (IS 2009 EC, IS 2010 PC, IS 2011 SSC) and SVM complexity $C \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$ was determined on the development set; the overall best result was achieved at $C = 0.02$ with the 2011 SSC feature set (cf. Fig. 2). The achieved UA and WA on the development and testing partitions of the SLC are shown in Table 7. In contrast to the Intoxication Sub-Challenge, the accuracy on the test set (70.3% UA) is considerably higher than the one on the development set (67.3% UA).

4.4. Challenge conditions

Having outlined the Challenge corpora and baseline results, we now describe the rules for the participants, their contributions and results.

As in the 2009 and 2010 Challenges, the labels of the test set were unknown, and all learning and optimisations needed to be based only on the training material. However, each participant could upload instance predictions to receive the confusion matrix and results up to five times. The upload format was instance and prediction, and optionally additional probabilities per class. This allowed a final fusion of all participants' results to demonstrate the potential maximum by combined efforts. As classes were unbalanced, the primary measure to optimise was unweighted accuracy (UA), i.e., unweighted average recall. The choice of unweighted average recall was a necessary step to better reflect imbalance of

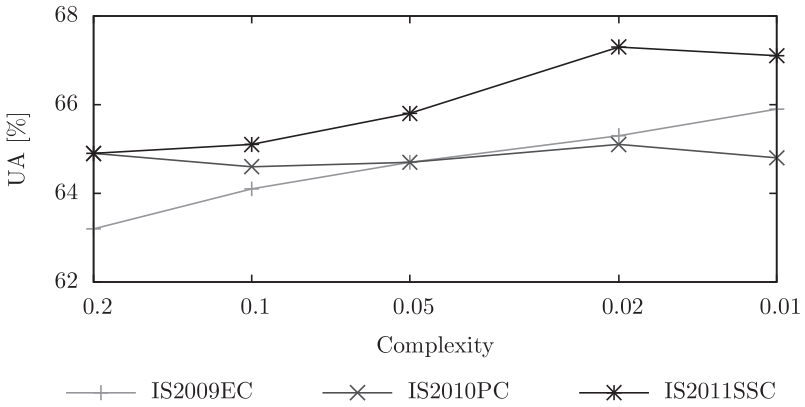


Fig. 2. Optimisation of SVM complexity by unweighted accuracy (UA) on the development partition of the SLC when training on the training partition after SMOTE. Three different feature sets are evaluated (cf. Table 7).

Table 7
Sleepiness Sub-Challenge baseline results by unweighted and weighted accuracy (UA/WA). SMO learned pairwise SVM with linear Kernel, complexity optimised on development partition to 0.02. SMOTE on (united) learning instances. Feature sets IS 2009 EC, IS 2010 PC, and IS 2011 SSC correspond to the official sets of the Challenges (Emotion, Schuller et al., 2009, Paralinguistic, Schuller et al., 2010a, and Speaker State, Schuller et al., 2011) held at INTERSPEECH in these years.

[%]	Train vs. Develop		Train + Develop vs. Test	
Features	UA	WA	UA	WA
IS 2009 EC	65.3	64.2	68.0	72.4
IS 2010 PC	65.1	66.4	70.2	72.8
IS 2011 SSC	67.3	69.1	**70.3**	72.9

The final baseline is set in bold face.

instances among classes as often given in real-world settings, where sober or alert speech is usually available in larger quantities than intoxicated or sleepy speech. Other well-suited and interesting measures such as the detection error trade-off were considered; however, these are either not yet common measures in the field or did not fit the evaluation paradigm, where participants are not required to submit classifier confidences.

While the competition measure was aimed at classification, for the training and development partitions of the ALC and SLC continuous valued annotations (BAC from 0.28 to 1.75 and mean KSS from 1 to 10) were provided. This information could be used for model construction or reporting of more precise results in submitted papers on the development partition. Furthermore, phonetic information was given for the ALC (cf. Section 4.2.1).

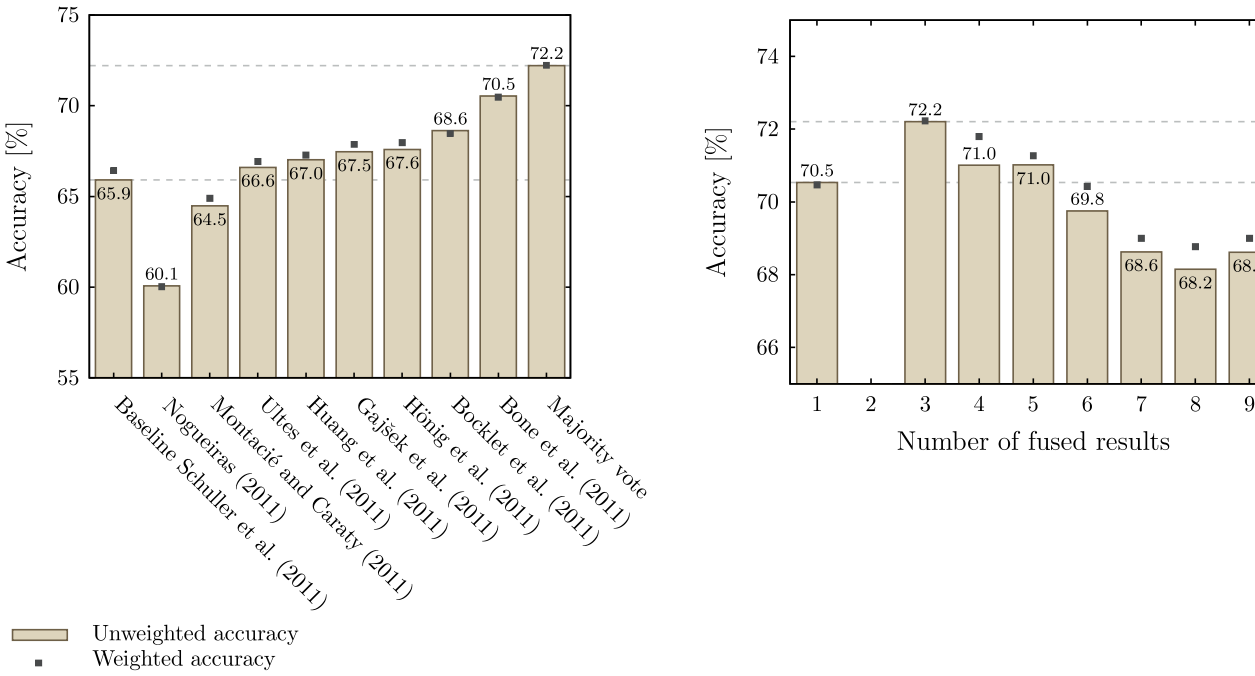
A new set of 4368 acoustic features per speech chunk, computed with TUM's openSMILE toolkit (Eyben et al., 2010) as in the 2009 and 2010 Challenges, was provided by the organisers. This set is based on applying functionals to 60 low level descriptors extracted on frame level (cf. above). These features could be used directly or sub-sampled, altered, or processed in any other way, and combined with other features. Both Sub-Challenges allowed contributors to find their own features with their own classification algorithm. The labels of the test set were unknown, and participants had to stick to the definition of training, development, and test sets. They were allowed to report on results obtained on the development set, but had only a limited number of five trials to upload their results on the test set, whose labels were unknown to them. Each participation had to be accompanied by a paper presenting the results that underwent peer-review. Only contributions with an accepted paper were eligible for Challenge participation. The organisers reserved the right to re-evaluate the findings, but did not participate themselves in the Challenge. Instead, they provided baselines using the standard WEKA toolkit (Hall et al., 2009) so that the results were reproducible. Participants were encouraged to compete in both Sub-Challenges.

4.5. Challenge results

All participants were encouraged to compete in both Sub-Challenges and each participant had to submit a paper to the INTERSPEECH 2011 Speaker State Challenge Special Event. Overall, 34 sites registered for the Challenge, and

Table 8
Participants of the intoxication and the Sleepiness Sub-Challenge.

Sub-Challenge		Participants	
Intoxication	Sleepiness	Count	Papers
✓		4	Ultes et al. (2011), Hönig et al. (2011), Bocklet et al. (2011) and Bone et al. (2011)
	✓	2	Bozkurt et al. (2011) and Rahman et al. (2011)
✓	✓	4	Nogueiras (2011), Montacié and Caraty (2011), Gajšek et al. (2011) and Huang et al. (2011)



(a) Results of the participants

(b) Majority vote of the best n systems

Fig. 3. Intoxication Sub-Challenge: (a) results of the participants and (b) majority vote of the best n systems.

12 papers were accepted for presentations. Ten groups actually submitted classification results on the official test set. Six groups took part in only one of the two Sub-Challenges: four groups in the Intoxication Sub-Challenge, and two groups in the Sleepiness Sub-Challenge. Four groups took part in both Sub-Challenges. Table 8 gives an overview of the participants and cites their contributions.

4.5.1. Contributions to the Intoxication Sub-Challenge

Fig. 3(a) summarises the results of the eight participants in the Intoxication Sub-Challenge on the official test set. Not all results surpassed the baseline result demonstrating the high competitiveness of the baseline. However, the proposed approaches are interesting, and in the official review process, they were considered worthwhile being published. Although the mid-range results are very close to each other, a large variety of different ways to address the classification problem can be observed.

In contrast to the other contributions, Nogueiras (2011) uses dynamic classification in order to model the temporal structure. Although the degree of intoxication is constant for the speech utterance under consideration, the speech changes over time. Thus, functionals computed over the whole utterance clearly depend on the phonetic contents. In order to cancel the effects of the phonetic contents, Nogueiras (2011) uses semi-continuous hidden Markov models with 32 states modelling the different phonemes. The HMM structure allows phonemes to occur in an arbitrary order. Based on a universal background model, individual HMMs are obtained for each of the two speaker states by re-estimation of the parameters using discriminative training. The 60 openSMILE low-level descriptors provided by the Challenge organisers with a context of 11 frames are used as acoustic features. Additionally, the mean feature vector for the

whole utterance is added. This high-dimensional feature space is reduced from 720 to 64 based on linear discriminant analysis (LDA). The results on the development set (63.0% UA) were close to the baseline result (65.3% UA on the development set). Unfortunately, only 60.1% UA were obtained on the test set.

In order to cancel the effects of the phoneme contents, Montacié and Caraty (2011) train gender-dependent SVM classifiers for each of the six phonemes `/a/`, `/6/`, `/aI/`, `/E/`, `/I/`, and `/l/`. As acoustic features, they use a subset of 30 features of the openSMILE features provided by the organisers, which they obtained by successively applying WEKA's feature selection algorithms *Subset Size Forward Selection* and *Best First*. The outputs of these six classifiers are combined using multi-layer perceptrons (MLP), support vector machines (SVM), and decision trees (J47 pruned trees). The latter method yields the best results. Additionally, the phoneme-based classifiers are combined with the baseline SVM classifier based on a selection of 30 features of the openSMILE Challenge feature set. Although the results on the development set are promising, the results on the test set (64.5%) are slightly below the official baseline (65.3%).

Ultes et al. (2011) enrich the official openSMILE feature set of 4368 acoustic features by adding 19 linguistic features based on the transcription: the total number of words per utterance and the number and rate of repetitions, hesitations, interruptions, corrections, word lengthenings, wrongly pronounced words, and pauses (additionally split into long and short pauses). As in Montacié and Caraty (2011), the set of features is reduced — yet using the information-gain ratio (IGR) — before classifying them with SVMs. The classification result on the test set is 66.6% UA and outperforms the baseline result.

Huang et al. (2011) rely on the official feature set of the INTERSPEECH 2011 Speaker State Challenge and the smaller feature sets of the two previous Challenges. For each of the three sets, they train a support vector machine. In order to cope with the fact that the two classes are unbalanced, they propose a classification technique called *Asymmetric SIMPLS*, which they use for all three data sets in addition to the SVM classifiers. Finally, the outputs of these six classifiers are combined using logistic regression, adaboost, and simple fusion. Best results are obtained with simple fusion, resulting in 67.0% UA on the test set.

Gajšek et al. (2011) use a UBM-MAP supervector approach based on the 26 acoustic features MFCC 1–12 and the RMS Energy, and their first order delta coefficients. Instead of training a Gaussian mixture model (GMM) as universal background model (UBM), the authors train 3-state left-to-right HMMs with one 26-dimensional Gaussian distribution in each state for each of the 47 allophones, based on the phone level transcriptions. MAP adaptation is used to adapt the mean vectors of the Gaussian distributions to a given utterance. Covariance matrices, weights, and transition probabilities are kept fixed. The transformed means form the supervector of dimension 47 [HMMs] · 3 [states] · 26 [dimensions] = 3666. These supervectors are classified with support vector machines. Further experiments are carried out with F0 based features. The two best systems are combined with the baseline system using a simple majority vote, yielding 67.5% UA.

Hönig et al. (2011) use 534 general-purpose prosodic features modelling durations, energy, pitch, and pauses for different segments (words, syllables, and nuclei). Furthermore, the authors use 17 'specialised' prosodic features: the duration of the whole chunk and the average duration of vocalic segments within one chunk, isochrony properties modelling the distance between consecutive stressed and consecutive unstressed syllables, variability indices modelling the differences in the duration of consecutive vocalic and consecutive consonantal segments, and global interval proportions (percentage of vocalic intervals, standard deviation of vocalic and consonantal segments). Features types are classified with support vector machines (SVM) and linear discriminant analysis (LDA). Finally, the output of the SVM classifier based on all prosodic features is combined with the output of the SVM classifier based on the openSMILE Challenge feature set. The authors identified a mismatch with respect to the spoken texts between Train + Develop and Test. After removing 30 prompts from Train + Develop, an unweighted accuracy of 67.6% was achieved on the test set, stating that the proposed features have an additional value under matched conditions.

Bocklet et al. (2011) developed several systems based on different feature sets and different classifiers. Gaussian mixture models (GMMs), which are obtained from an GMM universal background model after MAP adaptation, are used for spectral features. One system is trained on MFCCs, one on Perceptual Linear Prediction (PLP) features, and one on Temporal Patterns (TRAPS). Three more systems are obtained by classifying the same feature sets with a GMM supervector approach similar to Gajšek et al. (2011) except that GMMs are used instead of HMMs. System 7 is based on SVM classification of prosodic features computed on voiced speech segments. Another system is based on the official openSMILE feature set and SVMs. Further SVM systems use features based on the

transcription: the *phoneme duration system* uses duration statistics of pauses, schwas, vowels, and diphthongs, and mean and standard deviation features of phonemes and groups of phonemes. The *textual system* uses features such as the duration of the turn, the number of false, dialectal, or unintelligible words, the number of restarts, interrupts, irregularities, or hesitations, the approximate rate of speech, and a lexicality feature. Score level fusion is performed using linear logistic regression or simple majority vote. The best combination of systems achieves 68.6% UA on the test set.

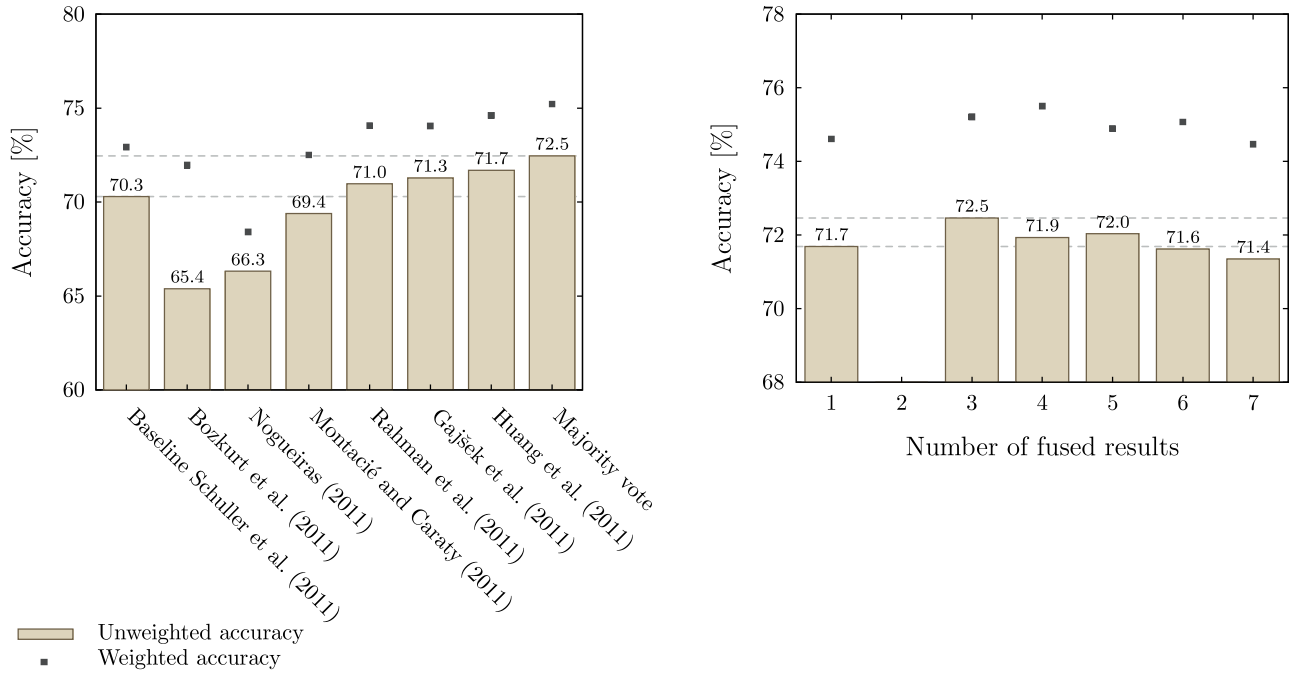
Bone et al. (2011) also use GMM supervector systems based on 39-dimensional MFCC features: one system is based on GMM mean supervectors, one on Tandem posterior probability supervectors, and the third one on Eigenchannel factor supervectors. Furthermore, the authors build one system using the official openSMILE feature set and one system based on the normalized pitch on a logarithmic scale, the normalised energy, and the first three formants and their bandwidths. These features are computed with Praat. In order to deal with widely varying utterance durations, the authors create another system with hierarchical features of the openSMILE and Praat low-level descriptors. First, they compute 15 functionals over windows of 0.1 s and 0.5 s, and finally they get features for the whole utterance by applying 6 'core' functionals on these features. Another system uses 103 global speech rate features computed from the phoneme durations that are extracted after forced alignment with the manual transcription. The authors use SVMs with a linear kernel. Instead of applying SMOTE, they use knowledge about the class bias to adjust the decision threshold of the SVM model. Bone et al. perform global and iterative speaker normalisation. The best model — 70.5% UA on the test set — was obtained by naïve fusion of all systems.

Biadsy et al. (2011) only report results on the development set and their own definition of a training and a test set. Unfortunately, results are also not given in terms of the unweighted accuracy, which is the official measure of the Challenge, making it hard to compare these results with the ones of the other participants. The authors build four systems modelling prosodic variations, phone duration variations, phonotactic variations, and spectral-phonetic variations. In order to model prosodic variations, prosodic events (pitch accents, intermediate and intonational phrase boundaries) are identified automatically using the AuToBI toolkit (Rosenberg, 2010), which is trained on Standard American English. *n*-Grams are used to represent the prosodic event sequence. Additionally, *n*-grams of deaccented words, the relative frequency of pitch accent, phrase accent and boundary tone types, and the overall accenting and phrasing rates, and the number of tones in the sequence are used. The classification with a logistic regression classifier shows poor results. According to the authors, this might be due to the short length of the utterances and the mismatch between the German speech of the database and the English speech the AuToBI toolkit is trained on. In order to model phone duration variations, statistics of phone durations extracted from a forced alignment of the transcription are computed and classified with logistic regression. Phonotactic variations are modelled with a bag-of-triphones approach using SVMs with a linear kernel. Spectral-phonetic variations are modelled with phone-dependent GMMs. For each of the 45 most frequent phones, a GMM universal background model is trained with 13 RASTA-PLP features (including energy) and their first and second order delta coefficients. MAP adaptation is used to adapt the GMM mean vectors to the realisations of the corresponding phone in a given utterance. SVMs with a KL-divergence-based kernel are used to classify these adapted phone-GMMs.

This summary of the Challenge contributions focuses on the automatic classification of sober and intoxicated speech. However, there are two contributions that also report results of human perception tests: Schiel (2011) and Ultes et al. (2011). These two papers are summarised in Section 2.4.

The given overview of the participants' contributions to the Challenge shows the large variety of different approaches. Still, the mid-range results are all very close. Fig. 5(a) shows which absolute improvements over a given experiment could be declared to be significantly better for the four levels of significance α =.050,.010,.005, and.001. The null hypothesis H0 assumes that the accuracies of both experiments are identical. We apply a one-tailed significance test since we are interested in whether the outcome of the second experiment is better than the first one. We assume that H0 is true and disprove it at various levels of significance. It depends on the accuracy of the first experiment which absolute improvements are necessary for the second one to be significantly better. Compared to the baseline (65.9%), accuracies $\geq$68.7 % could be considered to be significantly better at a level of $\alpha = .01$.

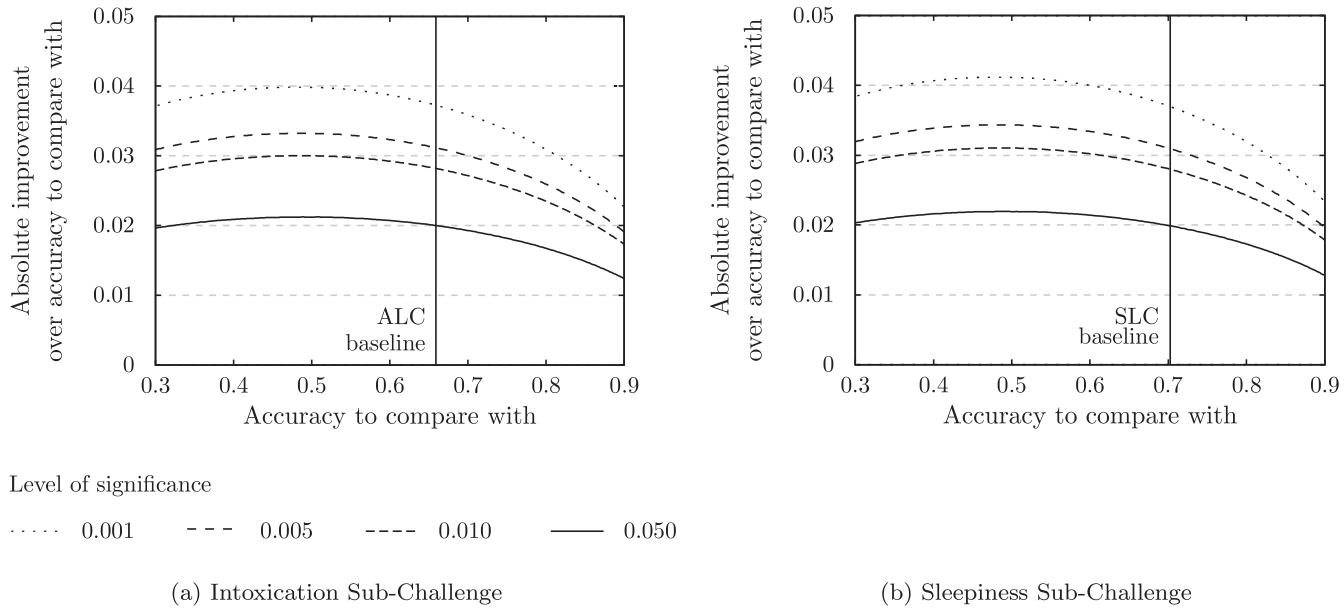
As in the previous Challenges, the individual results of the participants are combined by a simple majority vote of the best n participants. More sophisticated methods cannot be used since the participants' predictions are available only for the test set. Fig. 3(b) shows the results of this fusion for values of n between three and nine. If three to five systems are fused, the combined system outperforms the system of the winning team (70.5%). The best result (72.2%) is obtained for $n = 3$.



(a) Results of the participants

(b) Majority vote of the best n systems

Fig. 4. Sleepiness Sub-Challenge: (a) Results of the participants and (b) majority vote of the best n systems.



(a) Intoxication Sub-Challenge

(b) Sleepiness Sub-Challenge

Fig. 5. Significance of results: (a) Intoxication Sub-Challenge and (b) Sleepiness Sub-Challenge.

4.5.2. Contributions to the Sleepiness Sub-Challenge

Fig. 4(a) summarises the results of the six participants in the Sleepiness Sub-Challenge on the official test set. Three of the six results surpassed the baseline result (70.3% UA). Again, the baseline was highly competitive. Four of the six participants in the Sleepiness Sub-Challenges also took part in the Intoxication Sub-Challenge. Nogueiras (2011), Montacié and Caraty (2011), Gajšek et al. (2011), and Huang et al. (2011) applied the same system to both classification tasks. Their systems are described in the previous section.

As in the Intoxication Sub-Challenge, the performance of the system of Nogueiras Rodríguez (66.3% UA) and the one of Montacié and Caraty (69.4% UA) are below the baseline result.

The system of Gajšek et al. is based on a phoneme transcription, which is not available for the Sleepy Language Corpus (SLC). Therefore, they train a simple monophone recogniser on the Alcohol Language Corpus (ALC). Compared to the results in the Intoxication Sub-Challenge, their approach of modelling MFCCs shows a lower improvement probably due to inaccurate phoneme transcriptions, but it is still superior to the statistical approach and outperforms the baseline result: 71.3% UA vs. 70.3% UA. As in the Intoxication Sub-Challenge, the system of Huang et al. (71.7% UA) outperforms the baseline system in the Sleepiness Sub-Challenge, too, and is actually the highest result in this Sub-Challenge.

The contribution of Bozkurt et al. (2011) evaluates a training data selection method to prune possible outliers of mislabelled or ambiguous training samples. Their approach is based on Random Sampling Consensus (RANSAC). They use the official openSMILE feature set and SVMs with a linear kernel for classification. In contrast to the official baseline, no techniques such as SMOTE are used to handle the problem of unbalanced classes, resulting in clearly higher WA values than UA values. On the test set of the Sleepiness Sub-Challenge, the proposed approach (65.4% UA) outperforms the authors' own baseline (63.9% UA) but remains clearly below the official baseline in terms of the unweighted accuracy.

Rahman et al. (2011) evaluate feature level and decision level fusion of various systems. The baseline system is based on the official openSMILE features and SVMs with linear kernel and SMOTE. For the second system, a neutral GMM is trained on English speech of the Wall Street Journal-based Continuous Speech Recognition Corpus Phase II. A univariate GMM with four mixtures is trained for each feature of the baseline set. Then, the GMM likelihood scores are computed for sleepy and non-sleepy speech. Lower scores are expected for sleepy speech than for non-sleepy speech as sleepy speech is expected to deviate from neutral speech. SVMs are used for classification of these likelihood scores. The third system uses 17 features modelling local dynamics of the pitch contour. For each voiced segment within one utterance, functionals are applied to the F0 contour. In a second step, the mean, the maximum, and the standard deviation of these functionals are computed over all voiced segments of the same utterance. These features are then classified with SVMs along the same lines as the baseline system. The remaining two systems are GMM systems modelling 36-dimensional MFCCs (12 coefficients and their delta and delta-delta values), and 10-dimensional perceptual minimum variance distortionless response (PMVDR) features and their shifted delta cepstrum (SDC) features, respectively. For feature level fusion, all 7812 sentence-level features (baseline features, likelihood features, and F0 statistics) form one large feature vector, which is classified with SVMs. Additionally, the feature dimension is reduced using a chi-squared feature selection technique. However, no improvement over the baseline system is obtained. For decision level fusion, fusion with hard and soft decision labels is evaluated. The best result — 71.0% UA on the test set — is obtained with soft decision labels combining all single systems except the one based on F0 statistics.

As in the Intoxication Sub-Challenge, a large variety of different approaches can be observed in the Sleepiness Sub-Challenge. Again, many results are very close. Fig. 5(b) shows which absolute improvements over a given experiment could be declared to be significantly better for the four levels of significance $\alpha = .050, .010, .005$, and $.001$. Compared to the baseline (70.3%), accuracies $\geq$72.3 % can be considered to be significantly better at a level of $\alpha = .05$.

Again, the individual results of the participants are combined by a simple majority vote of the best n participants. Fig. 4(b) shows the results of this fusion for values of n between three and seven. If three to five systems are fused, the combined system outperforms the system of the winning team (71.7%). The best result (72.5%) is obtained again for $n = 3$.

4.5.3. Winners of the INTERSPEECH 2011 Speaker State Challenge

The results of the Challenge were presented in a Special Event of INTERSPEECH 2011 (double session) and the winners were awarded in the closing ceremony by the organisers. Two prizes (each 250.- GBP sponsored by the HUMAINE Association) could be awarded following the pre-conditions that the according paper needed to be accepted to the special event after the INTERSPEECH 2011 general peer-review, the provided baseline was exceeded, and a best result in a Sub-Challenge was reached.

The Intoxication Sub-Challenge Prize was awarded to Daniel Bone et al. (Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, Los Angeles) who reached 70.5% UA. The Sleepiness Sub-Challenge Prize was awarded to Dong-Yan Huang et al. (Institute for Infocomm Research/A*STAR, Singapore) who reached 71.7% UA.

Table 9
Pairwise Q-statistics measuring whether participants' systems commit the same errors on the test set. Avg: Average of Q-statistics with all other participants. The darker the shading, the higher the correlation.

(a) Intoxication Sub-Challenge

	Avg	#2	#3	#4	#5	#6	#7	#8	#9 (Nogueiras)
#1 (Bone et al.)	.427	.598	.425	.461	.460	.428	.433	.395	.217
#2 (Bocklet et al.)	.657		.710	.752	.757	.690	.740	.651	.356
#3 (Hönig et al.)	.754			.822	.941	.906	.946	.864	.418
#4 (Gajšek et al.)	.716				.836	.845	.880	.775	.358
#5 (Huang et al.)	.776					.917	.956	.885	.452
#6 (Ultes et al.)	.749						.967	.826	.417
#7 (Baseline)	.782							.903	.427
#8 (Montacié and Caraty)	.706								.346
#9 (Nogueiras)	.374								

(b) Sleepiness Sub-Challenge

	Avg	#2	#3	#4	#5	#6	#7 (Bozkurt et al.)
#1 (Huang et al.)	.940	.982	.951	.993	.977	.835	.904
#2 (Gajšek et al.)	.931		.934	.976	.958	.853	.882
#3 (Rahman et al.)	.899			.938	.906	.750	.913
#4 (Baseline)	.925				.958	.792	.893
#5 (Montacié and Caraty)	.912					.814	.862
#6 (Nogueiras)	.787						.676
#7 (Bozkurt et al.)	.855						

4.5.4. Analysing participants' systems: beyond accuracy-related measures

As stated above, evaluation of participant's contributions by accuracy reveals that systems are fairly close to each other in terms of overall performance. The question remains, though, whether all the systems fail on the same utterances (for example, because certain subjects successfully mask their state), or if they might have complementary strengths. We shed light on this aspect by considering the pairwise Q statistics (Yule, 1900; Afifi and Azen, 1979; Kuncheva and Whitaker, 2003). Informally, $Q_{A,B}$ measures whether two systems A and B commit the same errors on the evaluation set, information which is not contained in simple accuracy comparisons. More precisely,

$$Q_{A,B} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \tag{1}$$

where N^{11} and N^{00} are the numbers of instances where the predictions of A and B are both correct or incorrect, respectively, and N^{01} and N^{10} are the numbers of instances where only A or B commit an error.

For the Intoxication Sub-Challenge (Table 9a), we observe that the systems ranked first and last in the competition (Bone et al., 2011; Nogueiras, 2011) display low Q statistics with the other systems in the field. Indeed, this can be attributed to little methodological overlap: The system by Nogueiras (2011) is the only one relying on dynamic classification (Hidden Markov Models, HMM); the contribution by Bone et al. (2011) is the only one exploiting speaker normalisation techniques. Overall, the Q statistics seem to be highest among the middle-ranked participants. For the Sleepiness Sub-Challenge (Table 9b), differences in the Q statistics are much lower overall; yet again, the HMM system by Nogueiras (2011) displays the lowest Q values with the others.

Furthermore, the competition measure of the Intoxication Sub-Challenge, or related measures for binary classification or detection, do not reveal whether a system's accuracy in distinguishing alcoholised and non-alcoholised speech of a certain person depends on the person's actual intoxication level. In other words, even a system with high UA recall or equal error rate could still fail in recognising some of the most intoxicated speakers, due to the binarisation of the learning and prediction tasks. Hence, in Table 10a, we display the Spearman rank-correlation between the systems'

Table 10
Spearman's rank-correlation coefficients (ϱ) between (a) system accuracy per speaker and speaker BAC and (b) session-wise system accuracy and absolute KSS deviation from KSS threshold (7.5).

(a) Intoxication Sub-Challenge

Participant	ϱ
#1 (Bone et al.)	.493
#2 (Bocklet et al.)	.473
#3 (Hönig et al.)	.238
#4 (Gajšek et al.)	.197
#5 (Huang et al.)	.195
#6 (Ultes et al.)	.122
#7 (Baseline)	.118
#8 (Montacié and Caraty)	.110
#9 (Nogueiras)	.180

(b) Sleepiness Sub-Challenge

Participant	ϱ
#1 (Huang et al.)	.251
#2 (Gajšek et al.)	.276
#3 (Rahman et al.)	.358
#4 (Baseline)	.289
#5 (Caraty)	.232
#6 (Nogueiras)	.229
#7 (Bozkurt et al.)	.458

accuracy per speaker[5] and the speakers' BAC levels. A high correlation would mean that the system is more reliable in 'extreme cases' but less in 'limit cases'; conversely, a low correlation indicates that either some of the extreme cases are not recognised, or there is remarkable performance for some limit cases, or both. Given imperfect accuracy of automatic classification, the first kind of behaviour is arguably more desirable in practical applications. It turns out that only for the systems by Bone et al. (2011) and Bocklet et al. (2011) this correlation is significant ($p < .05$).

Following a similar procedure for the Sleepiness Sub-Challenge, in Table 10b we assess the correlation between the *session* accuracy (165 sessions where at least 10 utterances were recorded) and the absolute deviation of the KSS value of the session from the threshold value of 7.5.[6] The correlation is highest for the system by Bozkurt et al. (2011); this can probably be attributed to their automatic selection of 'prototypical' training instances.

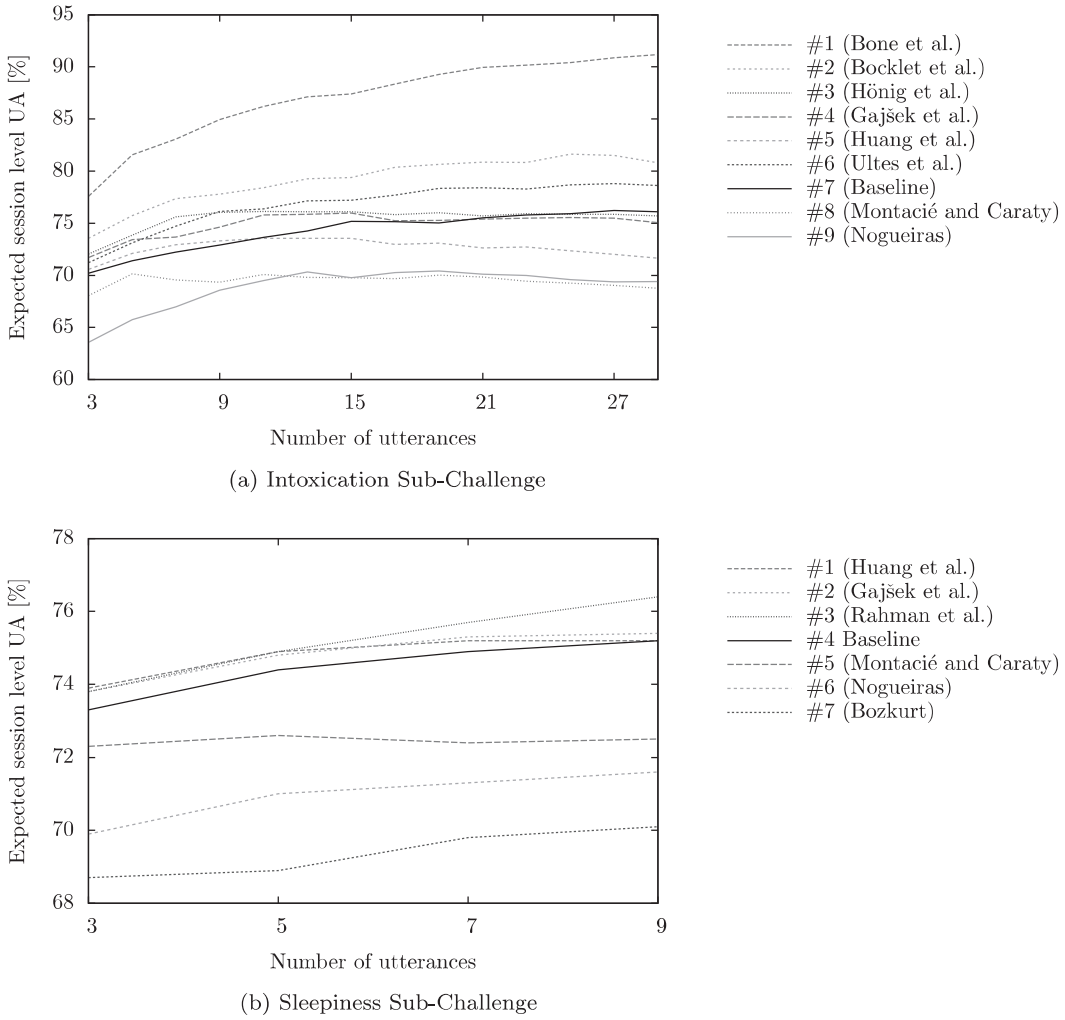
4.5.5. From short-term to medium-term: performance bounds for speaker state recognition

Overall, the results of both Sub-Challenges clearly demonstrate that utterance level recognition of intoxication and sleepiness is a demanding task. However, in many practical applications, medium term speaker states such as intoxication and sleepiness need not be recognised from single utterances; rather, observations from longer time intervals are available, in which the speaker state is supposed not to change. In the case of the Challenge data, we can consider entire recording sessions with one individual (of roughly 15 min in the ALC and 5–10 min in the SLC), during which intoxication and sleepiness level are assumed as constant. Hence, as an upper bound on what can be achieved with today's methodology, we consider decision level fusion of utterance level classifiers to gain a session level classification, as proposed by Weninger and Schuller (2011) for intoxication recognition. In that study, 76% UA recall of alcoholised/non-alcoholised recording sessions could be achieved on the Challenge test data by majority voting among the alcoholised/non-alcoholised utterance level decisions of the baseline classifier. We now proceed to apply this methodology to the results obtained by all of the Challenge participants, for both Sub-Challenges.

In particular, we investigate the relation between the number of utterances taken into account and the achieved accuracy to determine which amount of speech would be required in practice to achieve a robust decision. Precisely, we take the majority vote over N randomly selected utterances from each of the alcoholised (sleepy) and non-alcoholised (non-sleepy) sessions for each speaker. The parameter N is chosen from $\{3, 5, 7, \ldots, 29\}$ for the ALC and $\{3, 5, 7, 9\}$ for the SLC (odd numbers ensure that the majority vote is well-defined). Note that in the SLC, less speech material is available per session; further, we exclude all sessions from the SLC where less than nine utterances have been recorded. In the end, we consider all 100 sessions of the ALC test set and 189 of the SLC test set for the following experiments. For each value of N, the experiment is repeated 30 times with different random initialisations to deal with singular

[5] For the measurement of speaker specific system performance, we have to resort to conventional accuracy instead of UA, since not all speakers delivered alcoholised speech above.5 per mill BAC, hence the recall of the AL class per speaker is not well defined.

[6] The evaluation has to be performed slightly differently than for ALC, since in the SLC, there are multiple KSS values per speaker (yet only one per session), and naturally, there is no equivalent to a 'sober' condition.



(a) Intoxication Sub-Challenge

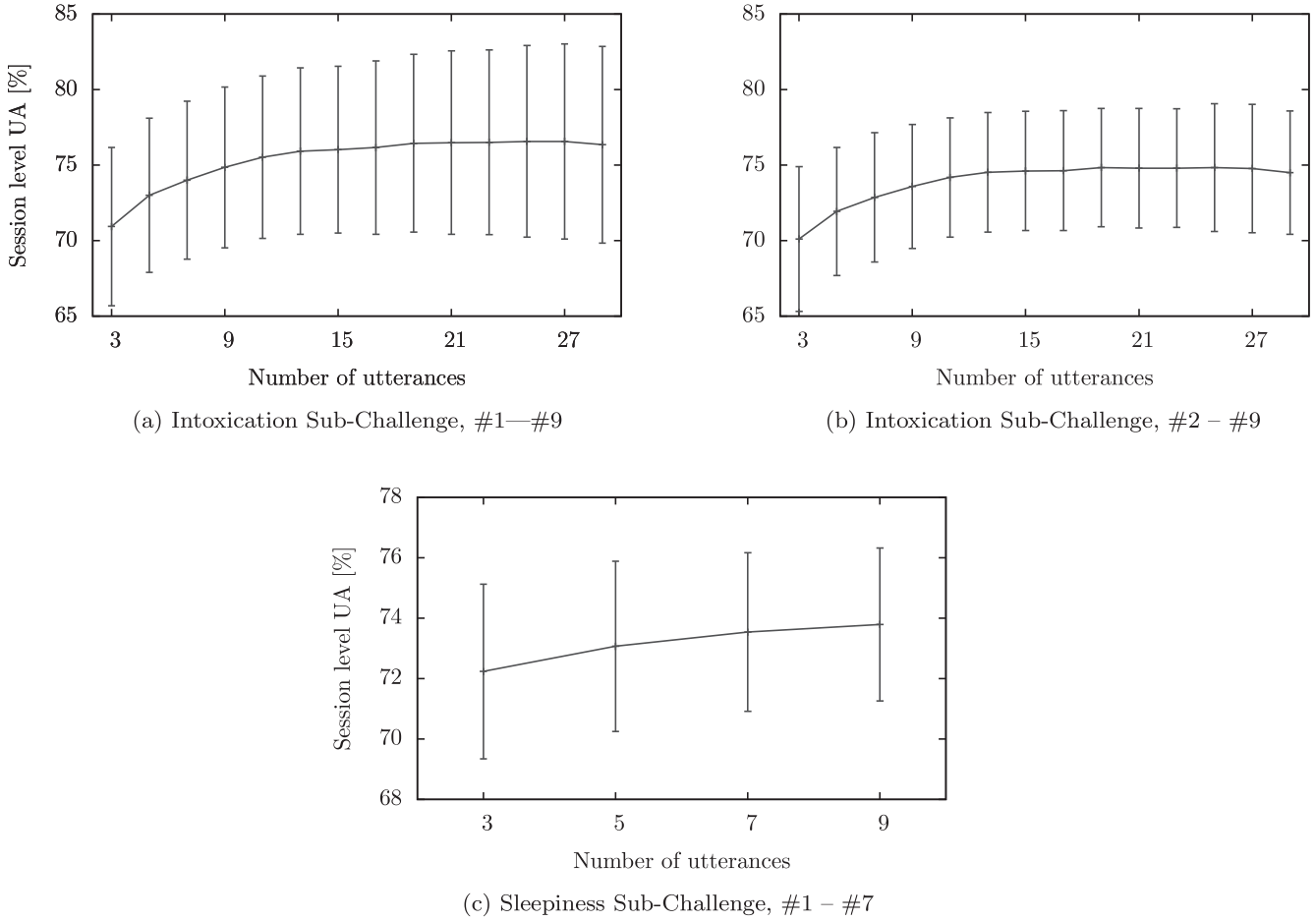
(b) Sleepiness Sub-Challenge

Fig. 6. Recognition of intoxication (a) and sleepiness (b) in a recording session through fusion of utterance level decisions by participants' systems. Expected UA on session level for increasing numbers of randomly selected utterances from each of the recording sessions. (a) Intoxication Sub-Challenge (b) Sleepiness Sub-Challenge.

effects due to 'lucky' or 'unlucky' selections; for each N, the expected UA (averaged across random initialisations) is reported in Fig. 6.

In the result, the best single system in the Intoxication Sub-Challenge (Bone et al., 2011) delivers a very remarkable average UA of 91% when voting over 29 utterances (Fig. 6(a)). The system by Bocklet et al. (2011) is ranked second with up to 81.6% average UA (for 25 utterances); interestingly, the system by Ultes et al. (2011), which is only slightly above the baseline and is ranked sixth on utterance level, is the third best on session level (up to 78.8% UA). Generally, the higher ranked systems can profit more from inclusion of more utterances in the majority vote; the systems by Huang et al. (2011) and Montacié and Caraty (2011) profit least, with the former even degrading performance with higher numbers of utterances taken into account. In order to investigate whether majority voting is beneficial on average, we calculate the average session accuracy across all participants and random seeds, as displayed in Fig. 7. Due to the system by Bone et al. (2011) being an outlier as visible in Fig. 6(a), we present a separate average across all participants except Bone et al. (2011) in Fig. 7(b). For both, the Intoxication (Fig. 7(a) and (b)) and Sleepiness Sub-Challenges (Fig. 7(c)) a somewhat upward trend is visible despite the large standard deviation which is caused on the one hand by variation among participants (as clearly visible in Fig. 6(a)), but also by the choice of utterances, since performance of automatic intoxication classification seems to depend heavily on the prompt type (spontaneous, read, or command and control speech) used for recording the utterances (Weninger and Schuller, 2011).

These remarkable differences in the accuracy of the majority voting, which exceed the differences in utterance level performance by far, clearly deserve some further investigation. Our hypothesis is that this is due to different



(a) Intoxication Sub-Challenge, #1—#9

(b) Intoxication Sub-Challenge, #2 – #9

(c) Sleepiness Sub-Challenge, #1 – #7

Fig. 7. (a) Recognition of intoxication (a), (b) and sleepiness (c) in a recording session through fusion of utterance level decisions by participants' systems. Mean and standard deviation of session level UA for increasing numbers of randomly selected utterances from each of the recording sessions, across various random seeds and all participants' systems (a), (c) or systems except Bone et al. (2011) (b), which is considered as an outlier. (a) Intoxication Sub-Challenge, #1–9 (b) Intoxication Sub-Challenge, #2–9 (c) Sleepiness Sub-Challenge, #1–7.

variability of the prediction accuracy per session: Obviously, since the number of utterances per session is constant in the Challenge test set, any utterance level accuracy above chance will result in perfect accuracy of the session level majority vote across all utterances, if the variability of the accuracy among sessions is zero. The hypothesis is confirmed by the significant negative correlation ($\varrho = -.75$, $p < .05$) between (i) the session level UA improvement by taking into account 29 instead of 3 utterances, and (ii) the standard deviation of the classification accuracy per session. Further, in considering for each system the correlations between its recalls of alcoholised and non-alcoholised utterances per speaker, we observe that for the systems except the one by Bone et al. (2011), a high recall of alcoholised utterances implies a low recall of non-alcoholised ones ($-.58 < \varrho < -.31$), while for the predictions by Bone et al. (2011), the recalls of alcoholised and non-alcoholised utterances per speaker are strongly correlated ($\varrho = .89$). This results in the observed great stability of the majority vote among utterances, while the other systems tend to be biased towards a decision for 'alcoholised' or 'non-alcoholised' for each speaker, and thus majority voting does not increase performance as drastically. This phenomenon can be attributed partly to the speaker normalisation as performed by Bone et al. (2011): Since the Challenge test set contains alcoholised and non-alcoholised utterances for each speaker, normalising such that each feature has unit variance and zero mean per speaker will contribute to maximising the separability of alcoholised and non-alcoholised utterances in the feature space, and minimise inter-speaker variation. However, note that speaker normalisation cannot be performed in an application scenario where only non-alcoholised or alcoholised speech is available for the speaker to be tested, e.g., at a police checkpoint.

Concerning the session level UA of majority voting among utterance level predictions from the Sleepiness Sub-Challenge (Fig. 6(b)), we observe differences to utterance level UA in the same order of magnitude as for the ALC, given the lower number of utterances that can be voted among. Notably, the system that performs best at session level

(Rahman et al., 2011; 75% UA) is not the best system at utterance level—in fact, it seems that the system of the Sub-Challenge winners (Huang et al., 2011) cannot exploit the increased amount of context, somewhat similarly to what we observe for their system in the Intoxication Sub-Challenge (Fig. 6(a)).

Going from here, a straightforward option would be voting over participants for each turn and combine the fused predictions over the utterances—however, in our experiments, this approach could never significantly exceed the performance of the single best system in majority voting. This indicates that there is much more information gained from observing multiple utterances than by fusing multiple systems' predictions.

5. Concluding remarks

The aim of this succession of three Interspeech challenges 2009, 2010, and 2011 has been two-fold: first, from a methodological point of view, we wanted to introduce the concept of a strict partition into train, development, and test, together with well-defined measures of performance — all this is known from established fields such as automatic Speech Recognition (ASR) — into the broad and divergent field of paralinguistics. Second, as for content-based research questions, we wanted to address different sub-fields of paralinguistics which we can describe, in somehow sloppy terms, as 'states and traits and all that is in-between, called medium-term'. In 2009 (Schuller et al., 2009), we addressed short-time emotional states such as 'anger' — a member of the established set of full-blown emotions — and a positive cover class consisting of 'joyful' as well as 'motherese', the latter definitely being no full-blown emotion but, at the same time, a well-defined interactional–emotional state whose description has a long tradition within developmental psychology. In 2010 (Schuller et al., 2010a), we dealt with pronounced speaker traits which we could describe as the 'primitives of personality', namely age and gender. Now, in this 2011 challenge, we addressed phenomena which are in between pronounced short-time states and long-time traits, namely intoxication and sleepiness. Both phenomena introduce on the one hand an interesting combination of annotations on the ordinal level (sleepiness) or measurements on the interval level (intoxication), both mapped onto a binary decision. On the other hand, the very essence of being 'medium-term' made it possible to have a look at decisions obtained for single units or 'accumulated' units, without any change of attribution. Comparing single systems and combined systems as well as different types of evaluation proved to be very instructive and might lead to a better understanding of differences between approaches and full systems.

All these states and traits are not only simply interesting phenomena; being able to deal with them, especially to obtain good classification performance, is a necessary prerequisite for incorporation into successful applications. And in turn, a further necessary prerequisite is to establish standards within these fields that make comparisons between studies and obtained performance possible. These standards include provision of feature sets that can be re-used as reference. We hope that this present challenge is a further step towards broadening the view and at the same time, defining and using standards within the field of paralinguistics.

The corpora used for the Challenge are available. The follow-up to the 2011 Speaker State Challenge is the 2012 Speaker Trait Challenge, which is on-going at the time of writing (Schuller et al., 2012). It focuses on "perceived" speaker traits — again for the first time in such a public and well-regulated comparative evaluation, featuring the tasks Personality, Likability, and Pathology.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 289021 (ASC-Inclusion), from the Bavarian Archive for Speech Signals and from the 'Bund gegen Alkohol und Drogen im Straßenverkehr' (BADS, 'Association against Alcohol and Drugs in Traffic'), and from the German Research Council under grant agreement KR 3698/4-1.

References

Afifi, A., Azen, S., 1979. Statistical Analysis. A Computer Oriented Approach. Academic Press, New York.

Ananthapadmanabha, T., 2011. Aerodynamic and acoustic theory of voice production. In: Neustein, A., Patil, H.A. (Eds.), Forensic Speaker Recognition, Law Enforcement and Counter-terrorism. Springer, New York, pp. 309–363.

Bard, E.G., Sotillo, C., Anderson, A.H., Thompson, H.S., Taylor, M.M., 1996. The DCIEM Map Task Corpus: spontaneous dialogue under sleep deprivation and drug treatment. Speech Communication 20, 71–84.

Batliner, A., Seppi, D., Steidl, S., Schuller, B., 2010. Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. Advances in Human–Computer Interaction 2010, 15, Article ID 782802.

Behne, D.M., Rivera, S.M., Pisoni, D.B., 1991. Effects of alcohol on speech: durations of isolated words, sentences and passages. Research on Speech Perception 17, 285–301.

Biadsy, F., Wang, W.Y., Rosenberg, A., Hirschberg, J., 2011. Intoxication detection using phonetic, phonotactic and prosodic cues. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3209–3212.

Bocklet, T., Riedhammer, K., Nöth, E., 2011. Drink and speak: on the automatic classification of alcohol intoxication by acoustic, prosodic and text-based features. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3213–3216.

Bone, D., Black, M.P., Li, M., Metallinou, A., Lee, S., Narayanan, S., 2011. Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3217–3220.

Bozkurt, E., Erzin, E., Eroğlu Erdem, C., Erdem, A.T., 2011. RANSAC-based training data selection for speaker state recognition. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3293–3296.

Bratzke, D., Rolke, B., Ulrich, R., Peters, M., 2007. Central slowing during the night. Psychological Science 18, 456–461.

Braun, A., 1991. Speaking while intoxicated: phonetic and forensic aspects. In: Proc. of International Congress of Phonetic Sciences (ICPhS), Aix-en-Provence, pp. 146–149.

Brenner, M., Cash, J., 1991. Speech analysis as an index of alcohol intoxication—the Exxon Valdez accident. Aviation. Space, and Environmental Medicine 62, 893–898.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357.

Chin, S.B., Pisoni, D.B., 1997. Alcohol and Speech. Academic Press Inc, San Diego, CA.

Cooney, O.M., McGuigan, K., Murphy, P., Conroy, R., 1998. Acoustic analysis of the effects of alcohol on the human voice. Journal of the Acoustical Society of America 103 (5), 2895.

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

Dinges, D., Kribbs, N., 1991. Performing while sleepy: effects of experimentally-induced sleepiness. In: Monk, T. (Ed.), Sleep, Sleepiness and Performance. Wiley, Chister, England, pp. 97–128.

Doddington, G.R., 1998. Sheep, goats, lambs and wolves—an analysis of individual differences in speaker recognition performance. In: Proc. of ICSLP, Sydney, Australia.

Echeburúa, E., de Medina, R.B., Aizpiri, J., 2007. Comorbidity of alcohol dependence and personality disorders: a comparative study. Alcohol & Alcoholism 42, 618–622.

Eyben, F., Wöllmer, M., Schuller, B., 2009. openEAR—introducing the munich open-source emotion and affect recognition toolkit. In: Proc. ACII, Amsterdam, pp. 576–581.

Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE—the Munich versatile and fast open-source audio feature extractor. In: Proc. ACM Multimedia, Florence, Italy, pp. 1459–1462.

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

Fulda, S., Popp, R., 2011. Measurement of daytime sleepiness in the elderly. Somnology 15, 154–159.

Gajšek, R., Dobrišek, S., Mihelič, F., 2011. University of Ljubljana system for interspeech 2011 speaker state challenge. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3297–3300.

Golz, M., Sommer, D., Mandic, D., 2005. Microsleep detection in electrophysiological signals. In: Dinesh, K., Hugo, G. (Eds.), Proc. of the 1st International Workshop on Biosignal Processing and Classification (BPC). Barcelona, Spain, pp. 102–109.

Golz, M., Sommer, D., Trutschel, U., Sirois, B., Edwards, D., 2010. Evaluation of fatigue monitoring technologies. Somnology 14, 187–189.

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

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., 2009. The WEKA data mining software: an update. SIGKDD Explorations 11 (1), 10–18.

Harrison, Y., Horne, J.A., 1997. Sleep deprivation affects speech. Sleep 20, 871–877.

Heinze, C., Trutschel, U., Schnupp, T., Sommer, D., Schenka, A., Krajewski, J., Golz, M., 2009. Operator fatigue estimation using heart rate measures. In: World Congress on Medical Physics and Biomedical Engineering, IFMBE Proceedings, Vol. 25, pp. 930–934.

Hollien, H., De Jong, G., Martin, C.A., Schwartz, R., Liljegren, K., 2001. Effects of ethanol intoxication on speech suprasegmentals. Journal of the Acoustical Society of America 110 (6), 3198–3206.

Hönig, F., Batliner, A., Nöth, E., 2011. Does it groove or does it stumble—automatic classification of alcoholic intoxication using prosodic features. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3225–3228.

Horberry, T., Hutchins, R., Tong, R., 2008. Motorcycle rider fatigue: a review. In: Road Safety Research Report, Vol. 78, Department for Transport, London, pp. 4–63.

Huang, D.-Y., Ge, S.S., Zhang, Z., 2011. Speaker state classification based on fusion of asymmetric SIMPLS and support vector machines. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3301–3304.

Ingre, M., Åkerstedt, T., Peters, B., Anund, A., Kecklund, G., Pickles, A., 2006. Subjective sleepiness and accident risk: avoiding the ecological fallacy. Journal of Sleep Research 15, 142–148.

Johannes, B., Salnitski, V.P., Gunga, H.-C., Kirsch, K., 2000. Voice stress monitoring in space—possibilities and limits. Aviation, Space, and Environmental Medicine 71, A58–65.

Johnson, K., Pisoni, D.B., Bernacki, R.H., 1990. Do voice recordings reveal whether a person is intoxicated? A case study. Phonetica 41, 215–237.

Kalivas, P.W., 2003. Predisposition to addiction: pharmacokinetics, pharmacodynamics, and brain circuitry. American Journal of Psychiatry 160, 1–3.

Klingholz, F., Penning, R., Liebhardt, E., 1988. Recognition of low-level alcohol intoxication from speech signal. Journal of the Acoustical Society of America 84 (3), 929–935.

Kostyk, B., Rochet, A., 1998. Laryngeal airway resistance in teachers with vocal fatigue: a preliminary study. Journal of Voice 12, 287–299.

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

Krajewski, J., Golz, M., Schnieder, S., Schnupp, T., Heinze, C., Sommer, D., 2010. Detecting fatigue from steering behaviour applying continuous wavelet transform. In: Proceedings Measuring Behaviour, Vol. 7, pp. 326–329.

Krajewski, J., Kroeger, B., 2007. Using prosodic and spectral characteristics for sleepiness detection. In: Proc. of Interspeech, Antwerp, pp. 1841–1844.

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

Kuncheva, L.I., Whitaker, C.J., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51 (2), 181–207.

Künzel, H.J., Braun, A., 2003. The effect of alcohol on speech prosody. In: Proc. of International Congress of Phonetic Sciences (ICPhS), Barcelona, pp. 2645–2648.

Künzel, H.J., Braun, A., Eysholdt, U., 1992. Einfluß von Alkohol auf Sprache und Stimme. Kriminalistik Verlag, Heidelberg.

Levelt, W.J.M., Roelofs, A., Meyer, A.S., 1999. A theory of lexical access in speech production. Journal of Behavioral and Brain Sciences 22, 1–75.

Levit, M., Huber, R., Batliner, A., Noeth, E., 2001. Use of prosodic speech characteristics for automated detection of alcohol intoxination. In: Bacchiani, M., Hirschberg, J., Litman, D., Ostendorf, M. (Eds.), Proc. of Workshop on Prosody and Speech Recognition. Red Bank, NJ, USA, pp. 103–106.

Lieberman, P., Kanki, B.G., Protopapas, A., 1995. Speech production and cognitive decrements on Mount Everest. Aviation, Space, and Environmental Medicine 66, 857–864.

Loukas, A., Krull, J.L., Chassin, L., Carle, A.C., 2000. The relation of personality to alcohol abuse/dependence in a high-risk sample. Journal of Personality 68, 1153–1175.

Martin, C.S., Yuchtman, M., 1986. Using speech as an index of alcohol-intoxication. Research on Speech Perception 12, 413–426.

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

Montacié, C., Caraty, M.-J., 2011. Combining multiple phoneme-based classifiers with audio feature-based classifier for the detection of alcohol intoxication. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3205–3208.

Morris, G.O., Williams, H.L., Lubin, A., 1960. Misperception and disorientation during sleep deprivation. Archives of General Psychiatry 2, 247–252.

Nogueiras, A., 2011. An HMM-based approach to the INTERSPEECH 2011 speaker state challenge. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3289–3292.

Nwe, T.L., Li, H., Dong, M., 2006. Analysis and detection of speech under sleep deprivation. In: Proc. of Interspeech, pp. 17–21.

O'Shaughnessy, D., 2000. Speech Communications: Human and Machine. IEEE Press, New York, USA.

Rahman, T., Mariooryad, S., Keshavamurthy, S., Liu, G., Hansen, J.H.L., Busso, C., 2011. Detecting sleepiness by fusing classifiers trained with novel acoustic features. In: Proc. of INTERSPEECH, Florence, Italy, pp. 3285–3288.

Read, L., 2006. Road Safety Part 1: Alcohol, Drugs and Fatigue. Department for Transport, London, 1-12.

Revelle, W., Scherer, K., 2009. Personality and emotion. In: Oxford Companion to the Affective Sciences. Oxford University Press, Oxford, pp. 1–4.

Rosenberg, A., 2010. AuToBI—a tool for automatic ToBI annotation. In: Proc. of INTERSPEECH, Makuhari, Japan, pp. 146–149.

Ruiz, R., Plantin De Hugues, P., Legros, C., 2010. Advanced voice analysis of pilots to detect fatigue and sleep inertia. Acta Acustica United with Acustica 96, 567–579.

Schenka, C., Schnupp, T., Heinze, C., Krajewski, J., Golz, M., 2010. The compensatory tracking task: a pattern recognition based approach for classifying vigilance. In: Proceedings Measuring Behaviour, Vol. 7, pp. 470–472.

Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Communication 40, 227–256.

Schiel, F., 2011. Perception of alcoholic intoxication in speech. In: Proc. of Interspeech, Florence, Italy, pp. 3281–3284.

Schiel, F., Heinrich, C., 2009. Laying the foundation for in-car alcohol detection by speech. In: Proc. INTERSPEECH 2009, Brighton, UK, pp. 983–986.

Schiel, F., Heinrich, C., Barfüßer, S., 2012. Alcohol language corpus. Language Resources and Evaluation 46 (3), 503–521, http://dx.doi.org/10.1007/s10579-011-9139-y, Springer, Berlin-New York.

Schiel, F., Heinrich, C., Neumeyer, V., 2010. Rhythm and formant features for automatic alcohol detection. In: Proc. of Interspeech, Chiba, Japan, pp. 458–461.

Schleicher, R., Galley, N., Briest, S., Galley, L., 2008. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? Ergonomics 51, 982.

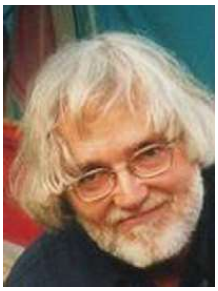
Schnieder, S., Krajewski, J., Esch, T., Baluch, B., Wilhelm, B., 2012. Just valid or even accurate: determine the measurement accuracy of the pupillographic sleepiness test by applying self- and observer ratings. Somnology, Sleep Research and Sleep Medicine 1, 1–15.

Schnupp, T.A., Edwards, S., Krajewski, D., Golz, J.M., 2009. Is posturography a candidate for a vigilance test? In: Proc. World Congress on Medical Physics and Biomedical Engineering, Vol. 25, pp. 388–392.

Schuller, B., Steidl, S., Batliner, A., 2009. The INTERSPEECH 2009 emotion challenge. In: Proc. of Interspeech, Brighton, UK, pp. 312–315.
Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2010a. The INTERSPEECH 2010 paralinguistic challenge—age, gender, and affect. In: Proc. of Interspeech, Makuhari, Japan, pp. 2794–2797.
Schuller, B., Steidl, S., Batliner, A., Nth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B., 2012. The INTERSPEECH 2012 speaker trait challenge. In: Proc. of Interspeech, Portland, OR.
Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., 2011. The Interspeech 2011 speaker state challenge. In: Proc. of Interspeech, Florence, Italy, pp. 3201–3204.
Schuller, B., Weninger, F., Wöllmer, M., Sun, Y., Rigoll, G., 2010 March. Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In: Proc. of ICASSP, Dallas, TX, USA, pp. 4562–4565.
Sforza, E., de Saint Hilaire, Z., Pelissolo, A., Rochat, T., Ibanez, V., 2002. Personality anxiety and mood traits in patients with sleep-related breathing disorders: effect of reduced daytime alertness. Sleep Medicine 3 (2), 139–145.
Shahidi, P., Southward, S.C., Ahmadian, M., 2010. Estimating crew alertness from speech. In: Proceedings of the American Society of Mechanical Engineers Joint Rail Conference, pp. 51–59.
Sigmund, M., Prokes, A., Zelinka, P., 2010. Detection of alcohol in speech signal using LF model. In: Proc. of International Conference on Artificial Intelligence and Applications, Innsbruck, pp. 193–196.
Sobell, L.C., Sobell, M.B., Coleman, R.F., 1982. Alcohol-inducted disfluency in non-alcoholics. Folia Phoniatrica 34, 316–323.
Sommer, D., Golz, M., Schnupp, T., Krajewski, J., Trutschel, U., Edwards, D., 2009. A measure of strong driver fatigue. In: Proc. of International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Vol. 4, pp. 9–15.
Story, B., 2002. An overview of the physiology, physics and modeling of the sound source for vowels. Acoustical Science and Technology 23, 195–206.
Traunmüller, H., 2000. Evidence for demodulation in speech perception. In: Proc. of ICSLP, Beijing, China, pp. 790–793.
Trojan, F., Kryspin-Exner, K., 1968. The decay of articulation under the influence of alcohol and paraldehyde. Folia Phoniatrica 20, 217–238.
Ultes, S., Schmitt, A., Minker, W., 2011. Attention, sobriety checkpoint! Can humans determine by means of voice, if someone is drunk. . . and can automatic classifiers compete? In: Proc. of INTERSPEECH, Florence, Italy, pp. 3221–3224.
Vogel, A.P., Fletcher, J., Maruff, P., 2010. Acoustic analysis of the effects of sustained wakefulness on speech. Journal of the Acoustical Society of America 128, 3747–3756.
Vollrath, M., 1993. Mikropausen im Sprechen. Peter Lang, Frankfurt, Germany.
Watanabe, H., Shin, T., Matsuo, H., Okuno, F., Tsuji, T., Matsuoka, M., Fukaura, J., Matsunaga, H., 1994. Studies on vocal fold injection and changes in pitch associated with alcohol intake. Journal of Voice, 340–346.
Weninger, F., Schuller, B., 2011. Fusing utterance-level classifiers for robust intoxication recognition from speech. In: Proc. Workshop on Inferring Cognitive and Emotional States from Multimodal Measures (MMCogEmS) held in conjunction with the 13th International Conference on Multimodal Interaction (ICMI 2011), Alicante, Spain.
Whitmore, J., Fisher, S., 1996. Speech during sustained operations. Speech Communication 20, 55–70.
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.
Yule, G., 1900. On the association of attributes in statistics. Philosophical Transactions of the Royal Society A 194, 257–319.
Zhang, X.-J., Gu, J.H., Tao, Z., 2010. Research of detecting fatigue from speech by PNN. In: Proc. of 2010 International Conference on Information Networking and Automation (ICINA). Vol. 2, pp. 278–281.

Björn Schuller received his diploma (1999), doctoral degree (2006), and habilitation (2012) all in electrical engineering and information technology from TUM in Munich/Germany where he is tenured as Senior Lecturer in Signal Processing and Machine Intelligence. He is currently with JOANNEUM RESEARCH in Graz/Austria, and was with the CNRS-LIMSI in Orsay/France (2009–2010), and visiting scientist in the Imperial College in London/UK (2010). Dr. Schuller is president-elect of the HUMAINE Association and member of the ACM, IEEE, and ISCA and (co-)authored more than 300 peer reviewed publications (2900 citations, h-index 28).

Stefan Steidl received his diploma degree in Computer Science in 2002 from Friedrich-Alexander University Erlangen-Nuremberg in Germany (FAU). In 2009, he received his doctoral degree from FAU for his work on Vocal Emotion Recognition. He is currently a member of the research staff of ICSI in Berkeley/USA and the Pattern Recognition Lab of FAU. His primary research interests are the classification of naturally occurring emotion-related states and of atypical speech (children's speech, speech of elderly people, pathological voices). He has (co-)authored more than 40 publications in journals and peer reviewed conference proceedings and been a member of the Network-of-Excellence HUMAINE.



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 Institute for Pattern Recognition 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 32 and more than 3500 citations. His research interests are all aspects of prosody and paralinguistics in speech processing. Dr. Batliner repeatedly served as Workshop/Session (co)-organiser and is Associated Editor for the IEEE Transactions on Affective Computing.

Florian Schiel received his Dipl.-Ing. and Dr.-Ing. degrees from TUM in Munich/Germany in 1990 and 1993 respectively, both in electrical engineering. Since 1993 he was mainly affiliated to the Institute of Phonetics at LMU in Munich/Germany, leading the VERBMOBIL, SmartKom, BITS and SmartWeb project groups. In 1994 and 1997 he was a research fellow at ICSI in Berkeley/California. In 2001 he earned the German 'Habilitation' at LMU and since then holds the chair of Phonetic Speech Processing. Currently he is CEO for BAS Services and is tenured as a senior researcher at the new Institute of Phonetics and Speech Processing at LMU.

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 University of Wuppertal and RWTH Aachen in Germany. He is Associate Professor in Industrial and Organizational Psychology in Würzburg since 2012 and vice director of the Center of Interdisciplinary Speech Science at the University of Wuppertal. He is member of the ISCA, Human Factors and Ergonomics Society, German Society of Psychology, and (co-)authored more than 50 publications in peer reviewed books, journals, and conference proceedings in the field of sleepiness detection, and signal processing.

Felix Weninger received his diploma in computer science (Dipl.-Inf. degree) from TUM in 2009. He is currently pursuing his PhD degree as a researcher in the Intelligent Audio Analysis Group at TUM's Institute for Human–Machine Communication, focusing on robust front-ends for speech and music information retrieval. He (co-)authored more than 40 publications in peer-reviewed books, journals and conference proceedings covering the fields of multi-source audio analysis, computational paralinguistics and medical informatics.

Florian Eyben obtained his diploma in Information Technology from TUM. He is currently pursuing his PhD degree in the Intelligent Audio Analysis Group. His research interests include large scale hierarchical audio feature extraction and evaluation, automatic emotion recognition from the speech signal, recognition of non-linguistic vocalisations, automatic large vocabulary continuous speech recognition, statistical and context-dependent language models, and Music Information Retrieval. He has over 70 publications in peer-reviewed books, journals and conference proceedings covering many of his areas of research, leading to over 700 citations and an H-index of 15.