The INTERSPEECH 2011 Speaker State Challenge*

Björn Schuller[1], *Stefan Steidl*[2,3], *Anton Batliner*[3], *Florian Schiel*[4], *Jarek Krajewski*[5]

[1]Technische Universität München, Institute for Human-Machine Communication, Germany
[2]ICSI, Berkeley, CA, USA
[3]FAU Erlangen-Nuremberg, Pattern Recognition Lab, Germany
[4]BAS, Ludwig-Maximilians-Universität München, Germany
[5]University of Wuppertal, Experimental Industrial Psychology, Germany

schuller@tum.de, steidl@icsi.berkeley.edu, batliner@informatik.uni-erlangen.de
schiel@phonetik.uni-muenchen.de, krajewsk@uni-wuppertal.de

Abstract

While the first open comparative challenges in the field of paralinguistics targeted more ‘conventional’ phenomena such as emotion, age, and gender, there still exists a multiplicity of not yet covered, but highly relevant speaker states and traits. The INTERSPEECH 2011 Speaker State Challenge thus addresses two new sub-challenges to overcome the usually low compatibility of results: In the Intoxication Sub-Challenge, alcoholisation of speakers has to be determined in two classes; in the Sleepiness Sub-Challenge, another two-class classification task has to be solved. This paper introduces the conditions, the Challenge corpora “Alcohol Language Corpus” and “Sleepy Language Corpus”, and a standard feature set that may be used. Further, baseline results are given.

Index Terms: Speaker State Challenge, Intoxication, Sleepiness

1. Introduction

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 broadens 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 [1]. For these Challenge tasks, the ALCOHOL LANGUAGE CORPUS (ALC) and the SLEEPY LANGUAGE CORPUS (SLC) with genuine intoxicated and sleepy speech are provided by the organisers. The first consists of 39 hours 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 hours of speech recordings of 99 subjects, annotated in the 10 different levels of sleepiness of the Karolinska Sleepiness Scale (KSS). The verbal material consists of different complexity reaching from sustained vowel phonation to natural communication. Partly, the corpora further feature detailed speaker meta data, orthographic transcript, phonemic transcript, segmentation, and multiple annotation tracks. Both are given with distinct definitions of test, development, and training partitions, with a strict speaker independence as needed in many real-life settings. Benchmark results are provided. In these respects, the INTERSPEECH 2011 Speaker State Challenge shall help bridging the gap between excellent research on paralinguistic information in spoken language and low compatibility of results. Two Sub-Challenges are addressed:

In the *Intoxication Sub-Challenge*, the alcoholisation of a speaker has to be determined as two-class classification task: *alcoholised* for a BAC exceeding 0.5 per mill or *non-alcoholised* for a BAC equal or below 0.5 per mill. The Challenge competition measure is the unweighted average recall (i. e., unweighted accuracy) of these two classes to better compensate for imbalance between classes. In the training and development partition, also the actual BAC from 0.28–1.75 per mill is provided. This information may be used as additional information for model construction or reporting of more precise results in submitted papers on the development partition.

In the *Sleepiness Sub-Challenge*, the sleepiness of a speaker has to be determined by a suited algorithm and acoustic features. While the annotation provides sleepiness from 1–10 by mean of annotations on the KSS, only two classes have to be recognised: sleepiness for a level exceeding level 7.5 on the KSS, and non-sleepiness for a level equal or below 7.5. Again, the full information on level of sleepiness is provided for the training and development partition, and the Challenge measure is unweighted average recall of the two classes.

Both Sub-Challenges allow contributors to find their own features with their own classification algorithm. However, a standard feature set is given per corpus that may be used. The labels of the test set are unknown, and participants will have to stick to the definition of training, development, and test sets. They may report on results obtained on the development set, but have only a limited number of five trials to upload their results on the test set, whose labels are unknown to them. Each participation needs to be accompanied by a paper presenting the results that undergoes peer-review. Only contributions with an accepted paper are eligible for Challenge participation. The organisers preserve the right to re-evaluate the findings, but do not participate themselves in the Challenge. Participants are encouraged to compete in both Sub-Challenges. We next introduce the Challenge corpora (Sec. 2), then features (Sec. 3), and baselines (Sec. 4), before concluding in Sec. 5.

* The authors would like to thank the sponsors of the challenge, the HUMAINE Association and Bavarian Archive for Speech Signals. The responsibility lies with the authors.

2. Challenge Corpora

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[1] please refer to [2, 3].

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 1.

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 blood alcohol concentration (BAC) he/she wanted to reach during the intoxication test. Using both Watson- and Widmark formula [3], the amount of required alcohol for each person was estimated and handed to the subject. After consumption, the speaker waited another 20 minutes before undergoing a breath alcohol concentration test (BRAC) 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[2]. Immediately after the tests, the speaker was asked to perform the ALC speech test which lasted no longer than 15 minutes, 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 minutes. 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 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. Three different speech styles are part of each ALC recording: read speech, spontaneous speech, and command & control. Speech styles are not marked for the Challenge.

2.2. 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, quantization 16 bit, microphone-to-mouth distance 0.3 m). The speech data consisted of different tasks: isolated vowels: sustained vowel phonation, sustained loud 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 a regular lecture. 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 1–10: 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 9 277 samples). A more detailed description of the data can be found in [4, 5].

For the Challenge, the available turns were 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 2.

Table 1: *Partitions of ALC. 'NAL' denotes recordings of non-alcoholized, i. .e., BAC per mill [0–0.5], and 'AL' recordings of alcolized speakers, i. e., BAC per mill]0.5–1.75].*

# ALC	**NAL**	**AL**	**total**
Train	3 750	1 650	5 400
Develop	2 790	1 170	3 960
Test	1 620	1 380	3 000
Train + Develop	6 540	2 820	9 360
Train + Develop + Test	8 160	4 200	12 360

Table 2: *Partitions of SLC. 'NSL' denotes recordings of non-sleepy, i. .e., KSS [1–7.5], and 'SL' recordings of sleepy speakers, i. e., KSS]7.5–10].*

# SLC	**NSL**	**SL**	**total**
Train	2 125	1 241	3 366
Develop	1 836	1 079	2 915
Test	1 957	851	2 808
Train + Develop	3 961	2 320	6 281
Train + Develop + Test	5 918	3 171	9 089

[1] The ALC corpus is available for unrestricted scientific and commercial usage. After the Challenge, interested parties may obtain copies of the full corpus at BAS (BAS distribution fees apply.). Please contact *bas@bas.uni-muenchen.de* or refer directly to the BAS catalogue at *www.bas.uni-muenchen.de/Bas*.

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

Table 3: *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-style 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
Spctral Flux, Entropy, Variance, Skewness, Kurtosis, Slope
5 voice related LLD
F0, Probability of voicing
Jitter (local, delta)
Shimmer (local)

3. Challenge Features

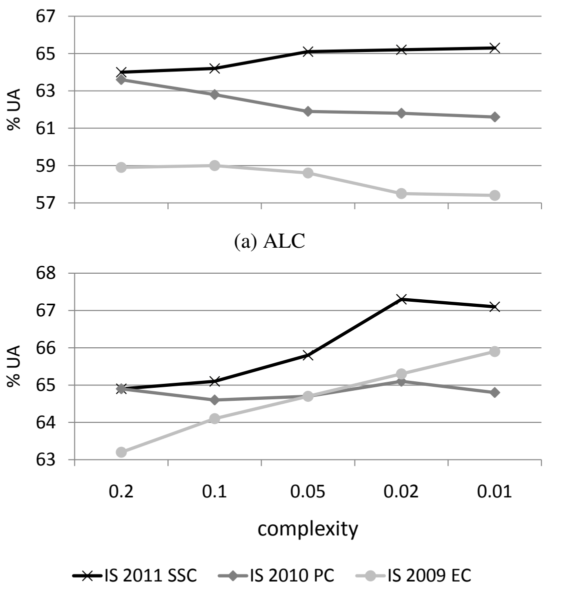
In this Challenge, an extended set of features with respect to the INTERSPEECH 2009 Emotion Challenge (384 features) [6] and INTERSPEECH 2010 Paralinguistic Challenge (1 582 features) [7] is given to the participants, again using the open-source Emotion and Affect Recognition (openEAR) [8] toolkit's feature extracting backend openSMILE [9]. The feature set consists of 4 368 features comprising features known as relevant for these tasks [10, 11] built from three sets of low-level descriptors and one corresponding set of functionals applied on the recording level for each LLD set. The LLD sets are given in Table 3: A major novelty concerning LLD compared to last year's challenge set 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.). Further, a base set of 33 functionals is introduced as shown in Table 4. Again, compared to last year's set the use of LPC coefficients and LP gain as functionals is new, as well as the standard deviation of the intra-peak distances. In the set of functionals applied to the spectral and energy related LLD, 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 $(\min + 0.25 \cdot \text{range})$ and $(\min + 0.75 \cdot \text{range})$. To the 54 energy and spectral LLD 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 3 996 features. To the 5 pitch and voice quality LLD 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 4. 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 alter and add features.

Table 4: *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

4. Challenge Baselines

For transparency and easy reproducibility, we use the WEKA data mining tool kit for classification [12], as we did for the INTERSPEECH 2009 Emotion Challenge and the 2010 Paralinguistic Challenge. As classifier we chose Support Vector Machines (SVM) with linear Kernel, Sequential Minimal Optimization (SMO) for learning, a linear Kernel function, and optimised the complexity on the development partition per corpus. Thereby, the complexity influences the number of Support Vectors for the hyperplane construction. We further use WEKA's implementation of the Synthetic Minority Over-sampling Technique (SMOTE) [13] as was done for the INTERSPEECH 2009 Emotion Challenge baseline, to balance instances in the respective 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 train partitions of ALC and SLC and testing on the respective development partitions is shown in Figure 1.a for ALC, and Figure 1.b for SLC in terms of unweighted accuracy – 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 all conditions on these tasks: Based on the optimal complexity as found on the development partitions, Table 5 shows baseline results for the *Intoxiation Sub-Challenge* (left) and *Sleepiness Sub-Challenge* (right) by unweighted and weighted accuracy on average per class (UA/WA, weighting with respect to number of instances per class). 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 can 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 can be uploaded five times by the participants.



(a) ALC

(b) SLC

Figure 1: *Unweighted accuracy (UA) of optimisation of SVM complexity on the development partitions of the ALC and SLC corpora when training on the training partitions after SMOTE. Three different feature sets are evaluated (cf. Table 5).*

5. Conclusions

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'. In 2009 [6], 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 [7], 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 address phenomena which are in between pronounced short-time states and long-time traits, namely intoxication and sleepiness.

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.

Table 5: *Intoxication and 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 0.01 (Intoxication Sub-Challenge) and 0.02 (Sleepiness Sub-Challenge). SMOTE on (united) learning instances. Feature sets IS 2009 EC, IS 2010 PC, and IS SSC 2011 correspond to the official sets of the Challenges (Emotion [6], Paralinguistic [7], and Speaker State) held at INTERSPEECH in the respective years.*

Sub-Challenge	***Intoxication***		***Sleepiness***	
Features	**% UA**	% WA	**% UA**	% WA
Train vs. Develop				
IS 2009 EC	57.4	65.3	65.3	64.2
IS 2010 PC	61.6	66.1	65.1	66.4
IS 2011 SSC	65.3	69.2	67.3	69.1
Train + Develop vs. Test				
IS 2009 EC	60.3	60.2	68.0	72.4
IS 2010 PC	63.2	62.6	70.2	72.8
IS 2011 SSC	**65.9**	66.4	**70.3**	72.9

6. References

[1] M. Brenner and J. Cash, "Speech analysis as an index of alcohol intoxication – the Exxon Valdez accident," *Aviation, Space, and Environmental Medicine*, vol. 62, pp. 893–898, 1991.

[2] F. Schiel and C. Heinrich, "Laying the Foundation for In-Car Alcohol Detection by Speech," in *Proc. INTERSPEECH 2009*, Brighton, UK, 2009, pp. 983–986.

[3] F. Schiel, C. Heinrich, and S. Barfüßer, "Alcohol Language Corpus," *Language Resources and Evaluation*, 2011, to appear.

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

[5] J. Krajewski and B. Kröger, "Using prosodic and spectral characteristics for sleepiness detection," in *Proc. INTERSPEECH 2007*, vol. 8, Antwerp, Belgium, 2007, pp. 1841–1844.

[6] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. INTERSPEECH 2009*, Brighton, UK, 2009, pp. 312–315.

[7] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. INTERSPEECH 2010*, Makuhari, Japan, 2010, pp. 2794–2797.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. ACII*, Amsterdam, 2009, pp. 576–581.

[9] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[10] S. B. Chin and D. B. Pisoni, *Alcohol and Speech.* Academic Press Inc, 1997.

[11] L. Dhupati, S. Kar, A. Rajaguru, and A. Routray, "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, 2010, pp. 917–921.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, 2009.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.