Improving Recognition of Speaker States and Traits by Cumulative Evidence: Intoxication, Sleepiness, Age and Gender

Felix Weninger, Erik Marchi, Björn Schuller

Institute for Human-Machine Communication, Technische Universität München, Germany
(weninger|marchi|schuller)@tum.de

Abstract

We address the fully automatic recognition of intoxication, sleepiness, age and gender from speech in medium-term observation intervals of up to several minutes. The nature of these speaker states and traits as being medium-term or long-term, as opposed to short-term states such as emotion, makes it possible to collect cumulative evidence in the form of utterance level decisions; we show that by fusing these decisions along the time axis, more and more accurate decisions can be obtained. In extensive test runs on three official INTERSPEECH Challenge corpora, we show that the average recall can be improved by up to 5 %, 6 %, 10 % and 11 % absolute by longer-term observation of speaker sleepiness, gender, intoxication, and age, respectively, compared to the accuracy of a decision from a single utterance.

1. Introduction

The emerging field of computational paralinguistics provides the chance to enhance technical systems with 'social competence', i. e., the ability to analyze and re-assess humans with respect to their traits (e. g., personality, age and gender) and states (e. g., affect, intoxication or sleepiness). Emulating such 'social skills' through signal processing and machine learning can be expected to be fruitful in multimedia retrieval, dialogue systems, but also for forensics and security. Particularly, recognition of (alcohol) intoxication and/or sleepiness from speech samples is an attractive non-invasive and pre-emptive method to improve security in high-risk environments such as driving, steering or controlling: Famous is the case of the Exxon Valdez accident in 1989 where phonetic analysis of the voice recordings clearly revealed the alcoholization of the ship's captain [1]. Furthermore, the increased usage of speech-based command and control interfaces in the car could be used to detect driver drousiness or sleepiness and warn her/him accordingly. In turn, main applications of automatic age and gender recognition are found in call classification and forensics.

Along the time axis, intoxication and sleepiness can be characterized as *medium-term* states, with a duration of at least several minutes (sleepiness) to usually a few hours (intoxication). Thus, similarly to long-term traits, they can be assumed to be constant during, e. g., interaction with an automated dialogue system, which clearly distinguishes them from short-term states such as emotion. This makes it possible to collect cumulative evidence for both, medium-term states and long-term traits, by observing an individual over the course of time and gradually refining the decision: In many applications there seems to be a tradeoff between the reliability of state and trait classification and the time required to obtain the decision, for example, to adapt dialogue strategies in an automated voice portal to a specific age/gender group, or to react to drousiness in a driver assistance system. Hence, in this study we propose to use an utterance level classifier, using a few seconds of speech material, and fuse its predictions across more and more utterances to increase reliability of the prediction over time. The benefit of this technique has been shown in [2] for alcohol intoxication; this paper aims to broaden this perspective by considering an additional state (sleepiness), as well as two well researched long term traits: age and gender [3, 4]. Consequently, in the case of medium-term states, we aim at a 'session level' classification (predicting the state of a human which is present in an interval of several minutes); conversely, for long-term traits, we predict a label for each speaker.

As another method to integrate knowledge from long-term observations, we evaluate speaker normalization for speaker state recognition, which has been proposed for the best performing system of the 2011 Speaker State Challenge (SSC) Intoxication Sub-Challenge [5]: If for each speaker, speech material from a sober (non-sleepy) condition is available in addition to recordings from an intoxicated (sleepy) condition, normalizing the feature values to zero mean and unit variance for each speaker (across both conditions) is expected to increase separability of both conditions in the feature space, and at the same time decrease the inter-speaker variance which results from invidual differences in the manifestation of intoxication and sleepiness in the speakers' voices.

Starting from this broad picture, the remainder of this paper details the evaluation databases (Section 2) taken from the series of INTERSPEECH Challenges in the field, the experimental setup (Section 3) and results (Section 4). Conclusions are drawn in Section 5.

2. Databases

2.1. Intoxication: Alcohol Language Corpus

We evaluate performance of intoxication recognition on the Alcohol Language Corpus (ALC) [6], the official corpus of the INTERSPEECH 2011 SSC, Intoxication Sub-Challenge [7]. The Challenge data set comprises a gender balanced subset of the ALC, 154 speakers (77 male, 77 female). The age range is 21–75 (mean: 31.0, standard deviation: 9.5 years). Speakers voluntarily underwent a systematic intoxication test where each speaker chose a blood alcohol concentration (BAC) and was handed the required alcohol. 20 minutes after consumption, the speaker underwent a BAC test and immediately afterwards, performed a speech test which lasted no longer than 15 minutes, to avoid significant changes caused by fatigue or saturation/decomposition of the measured BAC. : The BAC range in the ALC is between 0.28 and 1.75 per mill (volume of alcohol by volume of blood). At least two weeks later the speaker

Table 1: *Databases: Numbers of speakers, utterances and recording sessions (for speaker states).*

(a) Speaker States: Intoxication and Sleepiness. (N)AL: (non-)alcoholized; (N)SL: (non-)sleepy

	ALC				**SLC**			
	# spk.	# sessions (# utterances)			# spk.	# sessions (# utterances)		
		NAL	AL	total		NSL	SL	total
Train	60	65 (3 750)	55 (1 650)	120 (5 400)	36	249 (2 125)	111 (1 241)	360 (3 366)
Develop	44	49 (2 790)	39 (1 170)	88 (3 960)	30	224 (1 836)	80 (1 079)	304 (2 915)
Test	50	54 (1 620)	46 (1 380)	100 (3 000)	33	220 (1 957)	81 (851)	301 (2 808)
All	154	168 (8 160)	140 (4 200)	308 (12 360)	99	693 (5 918)	272 (3 171)	965 (9 089)

(b) Speaker Traits: Age and Gender. C = children, Y = young, A = adult, S = senior, F = female, M = male

	Agender: # speakers (# utterances)							
	C	YF	YM	AF	AM	SF	SM	total
Train	68 (4 406)	63 (4 638)	55 (4 019)	69 (4 573)	66 (4 417)	72 (4 924)	78 (5 549)	471 (32 526)
Develop	38 (2 396)	36 (2 722)	33 (2 170)	44 (3 361)	41 (2 512)	51 (3 561)	56 (3 826)	299 (20 548)

was required to undergo a second recording in sober condition, which took about 30 minutes. Additional details on the ALC can be found in [6]; interested parties may obtain copies of the full corpus from the Bavarian Archive for Speech Signals (BAS, www.bas.uni-muenchen.de). commercial usage. For the 2011 SSC, the continuous valued BAC was reduced to a binary label ('alcoholized' or 'non-alcoholized') based on a threshold of 0.5 per mill, which is the legal limit for driving in many countries. As a consequence, since not all participants reached a BAC over 0.5 per mill, the number of recording sessions classified as 'alcoholized' and 'non-alcoholized' is imbalanced (cf. Table 1a). Furthermore, we follow the speaker independent partitioning of the ALC as provided for the Challenge (cf. Table 1a).

2.2. Sleepiness: Sleepy Language Corpus

Second, sleepiness recognition is evaluated on the Sleepy Language Corpus (SLC), the official corpus of the 2011 SSC, Sleepiness Sub-Challenge [7]. 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 speech data consisted of read and spontaneous speech as detailed in [7]. A well established, standardized subjective sleepiness questionnaire measure, the Karolinska Sleepiness Scale (KSS), was used by the subjects (self-assessment) and additionally by three formally trained observers. In the version used in the present study, scores range from 1 (extremely alert) to 10 (cannot stay awake). For training and classification purposes, the recordings (mean KSS = 5.9, standard deviation = 2.2) were divided into two classes: non-sleepy ('NSL') and sleepy ('SL') with the threshold of 7.5, motivated by the high increase in accident risk above that threshold [8]. The average inter-observer agreement (Cohen's κ) for this binary decision is .814. Table 1a shows the speaker-independent Challenge partitioning of sessions and utterances.

2.3. Age and Gender: Agender Database

Third, as evaluation database for the recognition of speaker age and gender, we use the Agender database [9], the official corpus of the INTERSPEECH 2010 Paralinguistic Challenge (PC), Age and Gender Sub-Challenges [3]. An external company was employed to identify possible speakers of the targeted age and gender groups. Participants were asked to call an automated Interactive Voice Response system which prompted them to repeat given utterances or produce free content (e. g., the birth date of a family member). The number of calls, and hence, the exact number of utterances per speaker, varies in the corpus. For age classification, as in the Age Sub-Challenge of the 2010 PC, we consider the four-class problem to distinguish children (7–14 years), youth (15–24) years, adults (25–54 years) and seniors (55–80 years). This choice is mainly motivated from requirements of potential applications in automated voice portals, e. g., for marketing. The age distribution within these classes is uniform. All age groups, including the children, have equal gender distribution. Following the framework of the Gender Sub-Challenge of the 2010 PC, we treat gender classification as a three-way task (male, female, children), since gender discrimination of children is considerably difficult. The resulting number of speakers and utterances in the training and development sets of the Agender corpus is shown in Table 1b[1].

3. Experiments

Our experiments are evaluated in terms of unweighted average recall (UAR) of the classes, the official competition measure of the 2010 PC and 2011 SSC. The chance level UAR is 50 % for the binary intoxication and sleepiness recognition, 33 % for the three-way gender classification and 25 % for the four-class age recognition.

3.1. Utterance Level Classifiers

For age and gender classification on the utterance level, we use linear Support Vector Machines (SVMs) trained by the Sequential Minimal Optimization (SMO) algorithm as implemented in the Weka toolkit [10], as in the baselines of the 2010 PC [3]. Age and gender classification are carried out by joint learning of the seven possible age and gender groups (children and three age groups of two different genders) as proposed in [3], then mapping to the above named four age or three gender classes. Using the extended feature set of the INTERSPEECH 2012 Speaker Trait Challenge (STC) aiming at refined assessment of speaker traits [11], and a reduced complexity constant of 0.01 for the SVM training, we obtain 50.13 % UAR on the development set for age, which is clearly above the baseline of 47.11 % UAR presented in the 2010 PC baseline [3] ($p \ll 0.1$ % in a one-tailed z-test). For gender classification, we could also slightly improve the UAR (77.91 %) over the 2010 PC baseline

[1] The test set labels for Agender are not released at the time of this writing.

Table 2: *Influence of speaker normalization for recognition of medium-term speaker states: UAR on utterance level for binary classification of intoxication (ALC) and sleepiness (SLC).*

UAR [%]	**ALC**		**SLC**	
normalization	no [7]	yes	no [7]	yes
Train vs. Develop	65.3	**73.6**	**67.3**	57.0
Train + Develop vs. Test	65.9	**70.1**	**70.3**	59.1

(77.28 %).

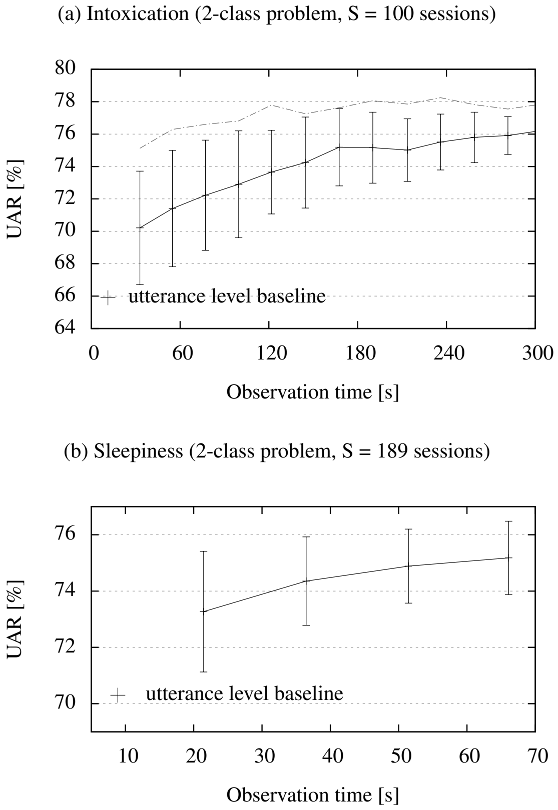
For intoxication and sleepiness classification, we exactly preserve the feature set fine-tuned to speaker state classification, and the classifier setup of the 2011 SSC [7]: We use SVMs with a complexity constant of 0.01 and 0.02 optimized on the ALC and SLC development sets, respectively. Speaker normalization is done for the training as well as the evaluation sets, without using class labels. The classifier training procedure is exactly the same with and without speaker normalization. We do not consider speaker normalization for speaker *trait* recognition: While inter-speaker variability in the feature space can be regarded as a confounding factor in speaker *state* recognition, it is obviously vital to enable speaker trait recognition.

3.2. Classification by Cumulative Evidence

For classification by 'cumulative evidence', i. e., by decision level fusion of utterance level classifiers over time, we are particularly interested in the relation between the observation time taken into account and the achieved accuracy to determine which amount of speech would be required in practice to achieve a robust decision. We believe that this is more meaningful than simply counting utterances—as done in our previous study on intoxication recognition [2]—, especially when comparing results across tasks.

For our experiments with speaker state recognition, we take the unweighted majority vote over N randomly selected utterances from each of the alcoholized (sleepy) and non-alcoholized (non-sleepy) sessions. The parameter N is chosen from $\{3, 5, 7, \ldots, 29\}$ for the ALC and $\{3, 5, 7, 9\}$ for the SLC (for binary classification, odd numbers ensure that the majority vote is well-defined). For age and gender, we perform a majority vote among the utterance level decisions for each speaker; here, we consider N in the range of $\{3, 4, \ldots, 30\}$, and in case of ties between two or more of the three or four classes, we randomly select a classification label. Note that while in the case of the ALC, all speakers have recorded at least 60 utterances, the number of utterances per session/speaker varies much more in the SLC and the Agender database. Thus, we exclude all sessions from the SLC where less than nine utterances have been recorded, as well as all speakers from the Agender database who provided less than 30 utterances. In the result, we consider all 100 sessions of the ALC test set and 189 of the SLC test set, as well as 271 speakers of the Agender development set for the following experiments. For each value of N, the experiment is repeated 30 times with different random number generator initialisations ('seeds') to deal with singular effects due to 'lucky' or 'unlucky' selections; for each N, the average total length (across seeds) of the considered utterances is plotted against the mean and standard deviation of the UAR (across seeds), in Figures 1 (intoxication and sleepiness) and 2 (age and gender). The 'utterance level baseline' point in these plots corresponds to the intersection of average utterance duration in the respective corpora and utterance level UAR.

Figure 1: *Classification of medium-term speaker states (intoxication and slepiness) by session level majority vote: Mean and standard deviation of UAR by average observation time for increasing numbers of randomly selected utterances. Dot-dashed line: Mean UAR with speaker normalization for intoxication recognition.*


(a) Intoxication (2-class problem, S = 100 sessions)

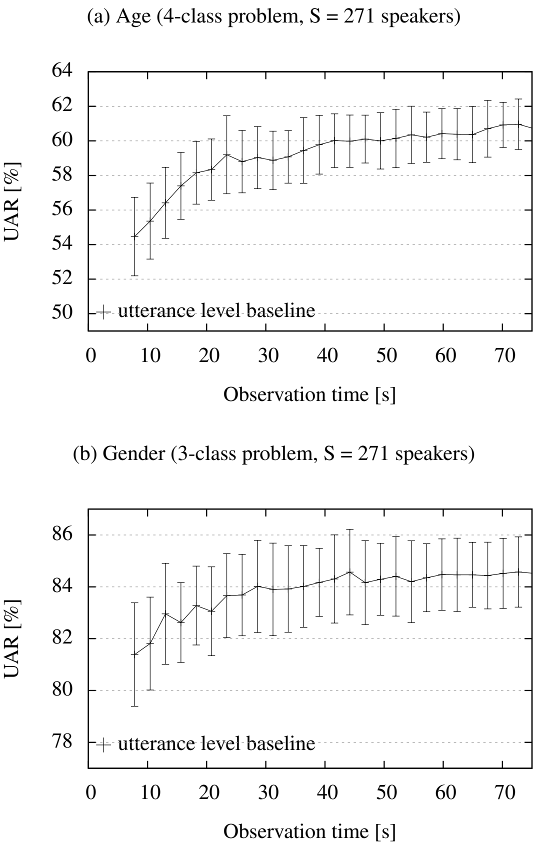
(b) Sleepiness (2-class problem, S = 189 sessions)

4. Results

4.1. Speaker Normalization

The influence of speaker normalization for speaker state recognition is shown in Table 2. We observe that for intoxication recognition, performance can be vastly improved by speaker normalization. On the development set, UAR is improved by over 8 % absolute; on the test set, we achieve 70.1 % UAR, which is more than 4 % absolute over the baseline and only slightly below the best performance in the 2011 SSC Intoxication Sub-Challenge (70.5 % UAR, [5])—notably, this system also uses speaker normalization, albeit in an iterative semi-supervised way. However, for the SLC, a vast decrease in performance is observed by speaker normalization (UAR of 59.1 %, which is 11 % absolute below the baseline on the test set). This can be attributed to both, the difference in the phenomena of intoxication and sleepiness, and the peculiarities of the corpora considered: In the ALC, the speaker states are much more 'controlled' by the fact that each speaker underwent one recording in a completely sober condition, while for the SLC, the range of participants' sleepiness levels varies considerably. In fact, there is no equivalent to 'completely sober' for sleepiness, and hence, no reference point that all of the non-sleepy samples can be mapped to. We further note that it is not clear whether the displayed performance improvement for intoxication recognition could be exploited in a real-life application

Figure 2: *Classification of long-term speaker traits (age and gender) by speaker level majority vote: Mean and standard deviation of UAR by average observation time for increasing numbers of randomly selected utterances.*


(a) Age (4-class problem, S = 271 speakers)

(b) Gender (3-class problem, S = 271 speakers)

where only speech from a single recording session (intoxicated or not) is available.

4.2. Cumulative Evidence

Next, we evaluate the performance of decision level fusion in Figures 1 (intoxication and sleepiness) and 2 (age and gender). To briefly summarize the results: For intoxication, over 76 % UAR on session level can be reached, as already reported in [2], when taking into account up to 5 minutes of speech; however, we observe a saturation in accuracy (75 % UAR already around 3 minutes of speech). We further observe a gain over the utterance level baseline (70 % UAR) already for half a minute of speech, albeit with a large standard deviation which can be partly attributed to the predictive power of different utterance types (read and spontaneous speech) in the ALC [2]. Combining cumulative evidence with speaker normalization, up to 78 % UAR can be reached. For sleepiness, a clear trend can be observed as well, leading up to 75 % UAR on the sessions of the test set for about a minute of observation time (utterance level baseline: 70.3 %). As expected, due to the inferior performance on utterance level, speaker normalization does not improve the results in the decision level fusion (max. 62 % UAR, not shown in the graph). For age and gender, largest improvements over the utterance level accuracy can be observed in the first thirty seconds, where UAR jumps over 59 % for age (utterance level: 50.1 %) and 84 % for gender (utterance level: 77.9 %). For age recognition, the UAR can be further improved to up to 61 %. Overall, the large improvements results in age and gender classification by fusion of a few short utterances suggest that maybe smaller chunks of data could be interesting for intoxication and sleepiness recognition as well.

5. Conclusions

In a large scale study on three official corpora from the INTERSPEECH 2010 and 2011 Challenges, we have shown that refinement of session and speaker level classification according to cumulative evidence is a very promising paradigm for the recognition of speaker states and traits. Future work should address an optimal unit of analysis for the utterance level decision to optimize the trade-off between the required observation time and the accuracy, and should include the modeling of ground truth changes within the observation interval, such as in system interaction with multiple speakers. Context modeling with recurrent neural networks could be a promising method for the latter.

6. Acknowledgment

The authors would like to thank the providers of the ALC, SLC and Agender database, Florian Schiel (Bavarian Archive for Speech Signals, Germany), Jarek Krajewski (University of Wuppertal, Germany) and Felix Burkhardt (Deutsche Telekom Laboratories, Germany). This study has received funding from the European Commission (grant no. 289021, ASC-Inclusion).

7. 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, no. 9, pp. 893–898, 1991.

[2] F. Weninger and B. Schuller, “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 ACM ICMI*, Alicante, Spain, 2011, no pagination.

[3] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, “Paralinguistics in Speech and Language - State-of-the-Art and the Challenge,” *Computer Speech and Language, Special Issue on Affective Speech in Real-Life Interactions*, 2012, 39 pages, DOI: 10.1016/j.csl.2012.02.005.

[4] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Müller, R. Huber, B. Andrassy, J. G. Bauer, and B. Littel, “Comparison of four approaches to age and gender recognition for telephone applications,” in *Proc. of ICASSP*, Honolulu, Hawaii, 2007, pp. 1089–1092.

[5] D. Bone, M. P. Black, M. Li, A. Metallinou, S. Lee, and S. Narayanan, “Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors,” in *Proc. of INTERSPEECH*, Florence, Italy, 2011, pp. 3217–3220.

[6] F. Schiel, C. Heinrich, and S. Barfüßer, “Alcohol Language Corpus,” *Language Resources and Evaluation*, 2011, DOI: 10.1007/s10579-011-9139-y.

[7] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, “The INTERSPEECH 2011 Speaker State Challenge,” in *Proc. of INTERSPEECH*, Florence, Italy, 2011, pp. 3201–3204.

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

[9] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, “A Database of Age and Gender Annotated Telephone Speech,” in *Proc. 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 2010, pp. 1562–1565.

[10] 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.

[11] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, “The INTERSPEECH 2012 Speaker Trait Challenge,” in *Proc. of INTERSPEECH*, Portland, OR, 2012.