Fusing Utterance-Level Classifiers for Robust Intoxication Recognition from Speech

Felix Weninger and Björn Schuller
Institute for Human-Machine Communication, Technische Universität München, Germany
{weninger|schuller}@tum.de

ABSTRACT

Obtaining speech samples is an attractive non-invasive method to recognize alcohol intoxication. In this paper, we aim to improve accuracy of speech-based intoxication recognition by decision fusion of utterance-level classifiers. On the official test set of the INTERSPEECH 2011 Speaker State Challenge (Intoxication Sub-Challenge), we demonstrate that up to 76.2 % accuracy can be obtained in binary classification on session level, which is 10 % absolute above the utterance level accuracy of the Challenge baseline.

1. INTRODUCTION

Relevant applications of automatic recognition of intoxication from speech are found in the medical domain and surveillance in high-risk environments such as driving, steering or controlling [1]. Hence, in the 'Intoxication Sub-Challenge' of the INTERSPEECH 2011 Speaker State Challenge [2] algorithms were evaluated on binary classification of speech utterances into 'non-alcoholized' (blood alcohol concentration (BAC) equal or below 0.5 per mill[1]) or 'alcoholized' instances (exceeding 0.5 per mill, which is the legal limit for car driving in Germany). Evaluation was carried out on the Alcohol Language Corpus (ALC) [3] with genuine intoxicated speech. In [2], benchmark results using brute forcing of acoustic features and Support Vector Machine (SVM) classification are given, achieving 65.9 % (unweighted) accuracy on the Challenge test set.

In this study, we investigate an arguably less challenging task on the same data which is motivated by practical application scenarios: The goal is to recognize from several utterances whether the speaker is above the legal BAC limit. To this end, we preserve the setup of the classifier training, the choice of purely acoustic features as well as the speaker-independent subdivision of the corpus from the Challenge. However, we fuse the prediction results of the classifier among multiple speech utterances of the same speaker in a single recording session of the corpus. 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. Furthermore, we investigate the influence of speech style on the fusion results.

[1]Per mill BAC by volume as standard in Germany and other European countries; resembles 0.05 per cent (Australia, Canada, USA).

MMCogEmS '11 Alicante, Spain

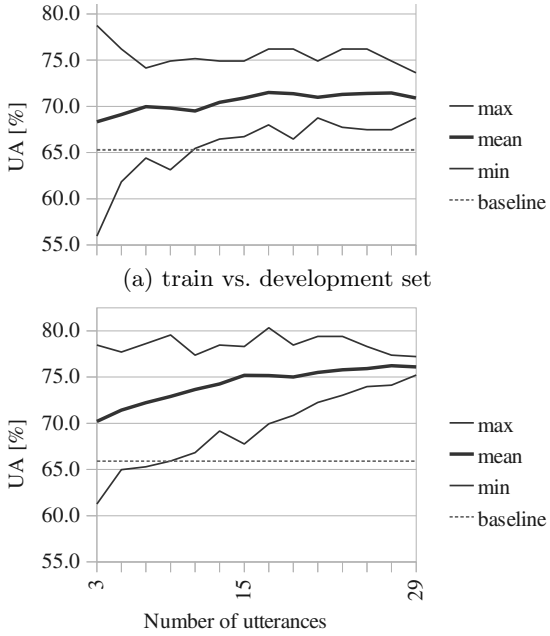
Table 1: Partitions of ALC. 'Spk': speakers. 'NAL': utterances with speaker BAC $\leq$ 0.5 per mill; 'AL': BAC $>$ 0.5 per mill.

#	Spk.	NAL	AL	total
Train	60	3 750	1 650	5 400
Devel	44	2 790	1 170	3 960
Test	50	1 620	1 380	3 000
Train + *Devel*	104	6 540	2 820	9 360
Train + *Devel* + *Test*	154	8 160	4 200	12 360

2. THE ALC CORPUS

The ALC corpus is available for unrestricted scientific and commercial usage from the Bavarian Archive of Speech Signals (BAS; distribution fees apply). In this study, the gender-based Challenge subset of 154 speakers (77 male, 77 female, age range 21–75) is selected for the experiments. Statistics on the training, development and test sets of the Challenge are shown in Table 1. To create the corpus, speakers voluntarily underwent a systematic intoxication procedure. Each speaker was handed the amount of alcohol to reach a self-chosen blood alcohol concentration (BAC). After consumption, each speaker underwent a blood sample test to determine the BAC. The BAC range in the corpus is between 0.28 and 1.75 per mill. Immediately after obtaining the BAC, each 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 session in sober condition, which took about 30 minutes. Thus, two sessions exist in the corpus for each speaker (one session after alcohol consumption, but not necessarily reaching a BAC of over 0.5 per mill, and one non-alcoholized session). All speakers are prompted with the same material. Three different speech styles are part of each recording session: read speech including 'tongue twisters', spontaneous speech, and command-and-control.



(a) train vs. development set

(b) train + development vs. test set

Figure 1: Session level unweighted accuracy (UA) on the ALC corpus by the number of (randomly selected) utterances for majority voting. Baseline: Utterance level accuracy from [2].

3. EXPERIMENTAL SETUP

We use the INTERSPEECH 2011 Speaker State Challenge baseline feature set comprising 4 368 acoustic features built from three sets of low-level descriptors (LLDs) known as relevant for intoxication detection [4] and one corresponding set of brute-forced functionals for each LLD set. This feature set has been shown to deliver higher classification accuracy than the lower-dimensional feature sets from the previous Challenges. The features are extracted using our open-source feature extraction toolkit openSMILE [5]. The classifier used in this study exactly corresponds to the setup of the Challenge baseline [2]. Linear SVM, trained with Sequential Minimal Optimization using a complexity constant of $C = 0.01$, are employed.

After classification, a majority vote is taken over N randomly selected utterances from each of the alcoholized and non-alcoholized sessions for each speaker. The parameter N is chosen from $\{3, 5, 7, \ldots, 29\}$ (odd numbers ensure that the majority vote is well-defined). The experiment is repeated 30 times with different random seeds to deal with singular effects due to 'lucky' or 'unlucky' selections. Mean, minimum and maximum unweighted accuracy (UA, average recall of the alcoholized / non-alcoholized classes) are reported over the 88 / 100 (development / test) sessions.

4. RESULTS AND DISCUSSION

The UA achieved by fusing the predictions on different numbers of utterances is shown in Figure 1. The utterance level baseline UA of 65.2 % / 65.9 % (development / test) roughly corresponds to the expected UA measured on session level when randomly picking a single utterance per session. Thus, the results indicate that the expected (mean)

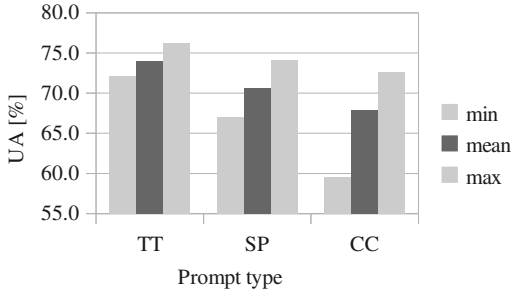


Figure 2: Session level unweighted accuracy (UA) on ALC development set by majority voting among five utterances of a single prompt type (TT: tongue twister; SP: spontaneous speech; CC: command-and-control).

UA can be constantly improved by majority voting among more and more utterance level decisions. Up to 71.4 % mean UA (on development) and 76.2 % (on test) are obtained; the optimum number of utterances is 27 for both sets; yet a random pick of as little as three utterances already improves the mean UA on the test set drastically to 70.2 % (4.3 % absolute improvement). To explain the high variation due to the random choice of utterances, we shed light on the influence of speech style: In Figure 2, the UA by majority voting among 5 utterances for three different prompt types in the corpus is displayed. It is evident that tongue-twisters exhibit the greatest robustness and smallest variation. Command-and-control utterances lag considerably behind on average, probably due to their simplicity; interestingly, their variation in performance is observed highest. Finally, spontaneous speech seems remarkably effective—remember that in this study, only acoustic features are employed; thus, the arguably higher variability of spontaneous speech in comparison to read speech under intoxication may be advantageous.

5. CONCLUSIONS

We have shown that the automatic intoxication recognition from speech can be made more robust by majority voting on the classifier decisions obtained on more than one speech utterance. Furthermore, considerable performance differences have been revealed concerning the usage of different prompt types for recording. Future work could focus on fusion with lexical features and automatic speech recognition confidence measures, as well as 'matched-condition' learning for the various prompt types.

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] 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.

[3] 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.

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

[5] 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.