Towards Automatic Intoxication Detection from Speech in Real-Life Acoustic Environments

Zixing Zhang, Felix Weninger and Björn Schuller

Institute for Human-Machine Communication, Technische Universität München, Germany
Email: `{zixing.zhang, weninger, schuller}@tum.de`
Web: `http://www.mmk.ei.tum.de`

Abstract

In-car intoxication detection from speech is a highly promising non-intrusive method to reduce the accident risk associated with drunk driving. However, in-car noise significantly influences the recognition performance and needs to be addressed in practical applications. In this paper, we investigate how seriously the intrinsic in-car noise and background music affect the accuracy of intoxication recognition. In extensive test runs using the official speech corpus of the INTERSPEECH 2011 Intoxication Challenge, realistic car noise and original popular music we conclude that stationary driving noise as well as music introduce a significant downgrade when acoustic models are trained on clean speech only, which can partly be alleviated by multi-condition training. Besides, exploiting cumulative evidence over time by late decision fusion appears to be a promising way to further enhance performance in noisy conditions.

1 Introduction

It is generally known that drunk driving is the cause for many traffic accidents worldwide. In addition to law enforcement, information technology can play an important role in solving this problem: In contrast to the legally binding blood and breath samples, non-invasive and non-intrusive monitoring of the driver can be used to recognize alcohol intoxication pre-emptively (i.e., before an accident has happened), and warn drivers accordingly. Technical systems for driver monitoring include face and gesture detection and identification [1], or physical sensors [2].

However, these often require expensive surveillance equipment that is not found in today's cars. In contrast, due to the increasing amount of in-car speech interfaces, detection of driver intoxication based on the speech signal to detect drivers' intoxication state becomes more and more feasible and interesting. In [3], the Alcohol Language Corpus (ALC) of genuine intoxicated speech is presented. In the INTERSPEECH 2011 Speaker State Challenge (SSC) [4], benchmark results for binary classification (below/above 0.5 per mill of blood alcohol concentration) are given, reaching up to 65.9 % unweighted average recall using brute forced acoustic features and Support Vector Machines (SVMs). These benchmark results, nevertheless, do not take into account a realistic car environment, which includes intrinsic noise (engine noise, road friction sounds, etc.) as well as non-intrinsic noise generated mainly by passengers and entertainment systems, such as playback of music.

In this paper, we address two questions: The first one is how seriously these noises affect the system. The second one is whether there are some methods to mitigate the noise influence and promote the robustness of intoxication detection system in car. In the field of automatic speech recognition (ASR), acoustic model adaptation is widely used for fitting various acoustic environments [5, 6]. However, to the best of our knowledge, such techniques have not been investigated yet for general speaker state recognition. In this paper, we rely on multi-condition training which is straightforward to integrate into discriminatively trained models such as SVMs and has delivered promising results for ASR tasks in non-stationary noise [7]. Additionally, we exploit the strategy proposed in [8] for collecting cumulative evidence in the form of utterance level decisions to gain a more robust classification of 'medium-term' speaker states such as intoxication. The crucial question is whether this method generalizes to the in-car acoustic environment.

In the following, Section 2 introduces three databases corresponding to intoxicated speech (ALC), in-car driving noise, and MTV music database. The selection of extracted feature sets and classifier follows in Subsection 3.1. After giving a brief overview of the experimental setup in Subsection 3.2, models adapted to multiple in-car acoustic environments are investigated in Subsection 3.3. Further, the classification by the 'cumulative evidence' strategy is evaluated in Subsection 3.4. Finally, Section 4 draws the conclusions.

2 Databases

2.1 Alcohol Language Corpus

The ALC [3] contains 38 hours of genuine alcohol intoxicated and sober speech. For our experiments, as for the 2011 SSC, we use a gender balanced subset of the ALC with 154 speakers (77 male, 77 female). Speakers are within the age range of 21 to 75 years and were selected to ensure a balance of German dialects. The corpus is subdivided into training, testing and development partitions guaranteeing speaker independence. Table 1 shows these partitions in detail.

Table 1: *Partitions of ALC. Spk.: number of speakers; 'NAL': number of recordings from speakers with BAC $\leqslant$ 0.5 per mill; 'AL': recordings from speakers with BAC $>$ 0.5 per mill*

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

To create the corpus, speakers were recorded at self-chosen blood alcohol concentrations (BACs) ranging from 0.28 to 1.75 per mill. The intoxicated speech material in the ALC was obtained by a speech test which the speakers were asked to perform immediately after taking a blood sample. Since the speech test did not last longer than 15 minutes, it is ensured that the BAC throughout the speech test remains roughly equal to the measured BAC before the test. At least two weeks after the intoxicated speech test, each speaker returned to undergo a second recording in sober condition. The sober recordings were chosen to be roughly twice as long as the intoxicated recordings.

Three different speech styles are part of each ALC recording: read speech, spontaneous speech, and command & control. The three partitions of the ALC corpus are mixed with additive driving noise as well as random segments of the MTV Music Database (cf. Subsection 2.3 and 3.2).

2.2 In-Car Driving Noise

To simulate realistic in-car intoxication recognition, we consider different types of interior car noises. In-car driving noises are mainly generated by wind, the engine, wheel friction, pounding or relative movement of car components, etc. The driving noise used in this paper was recorded in a MINI Cooper convertible[1] which presents a 'worst case scenario' regarding driving noise compared to other types of vehicles.

Since the road surface has strong influence on the characteristics of interior driving noise, we consider three kinds of road surfaces (smooth city road, highway, big cobble) with corresponding typical velocities. The lowest noise level is encountered when driving on a smooth city (CTY) road at 50 km/h; a medium noise level is measured for a highway (HWY) drive at 120 km/h; and the worst and loudest sound in interior of a car is provoked by a road with big cobbles (COB) at a velocity of 30 km/h.

2.3 MTV Music Database

As an example of realistic music that is likely to be encountered in the car, we use the MTV corpus of popular music. This corpus consists of 200 songs from the collection 'Twenty Years on MTV' covering the years from 1981 to 2000 as well as various genres from hip-hop to country music, and featuring male as well as female singers. To guarantee the independence of music added to the training, testing and development parts of ALC speech, and thus prevent overadaptation, the whole song set is divided into three sets with 100, 50 and 50 songs corresponding to the three partitions of ALC.

3 Experiments and Results

3.1 Feature Extraction and Classifier

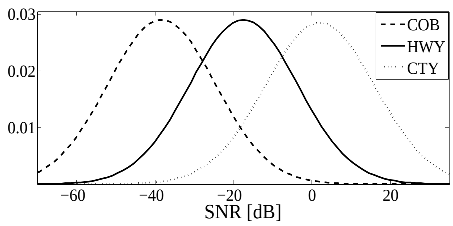
The acoustic feature vectors correspond to the INTERSPEECH 2011 Speaker State Challenge feature set [4] with 4 368 features generated by extracting 60 low-level descriptors (LLD) and applying 39 functionals, extracted by our feature extraction toolkit openSMILE. The set of 60 LLDs includes 4 energy related, 50 spectral related, 5 voice related LLDs, and the F0 contour. Further, the first order delta regression coefficients of the LLDs are computed. For the details of LLDs and functionals, please refer to [4]. As classifiers, SVMs with linear kernel, complexity of 0.05, and training by Sequential Minimal Optimization (SMO) are chosen in our experiments. The WEKA toolkit is used as in the challenge baseline.

3.2 Experimental Setup

In order to investigate the influence of in-car driving noises and musical sounds on drivers' intoxication state detection, three acoustic scenarios are considered: 1) clean ALC speech; 2) ALC speech overlaid by in-car driving noise; 3) ALC speech combined with in-car driving noise and background music.

For the second scenario, the original volume of the in-car intrinsic noise caused by driving is added to the ALC speech. Thus, three variants of the corpus are generated, each mixed with one of three different types of in-car driving noise (COB, HWY, CTY). Figure 1 displays the (fitted Gaussian) distributions of speech-to-noise ratios (SNRs) when driving on cobble, highway and city road surfaces. The mean SNRs for these three noise types are roughly -35, -15, and 5 dB, respectively. Note that despite these very high noise levels, the speech is still audible since the car noise is mostly limited to low frequency ranges. For the third scenario, we take into account four levels of speech-to-music ratio (SMR): 20, 15, 10, and 5 dB. Note that these ratios are calculated after adding driving noise, as it is likely that a driver would adjust the music volume according to velocity and surface. Therefore, twelve permutations of car noise and music noise levels exist. Overall, 16 (1+3+12) variants of the ALC are obtained corresponding to the scenarios 1–3, respectively.

Figure 1: Gaussian distribution of SNRs when driving on three road surfaces: cobble (COB), highway (HWY), city (CTY)



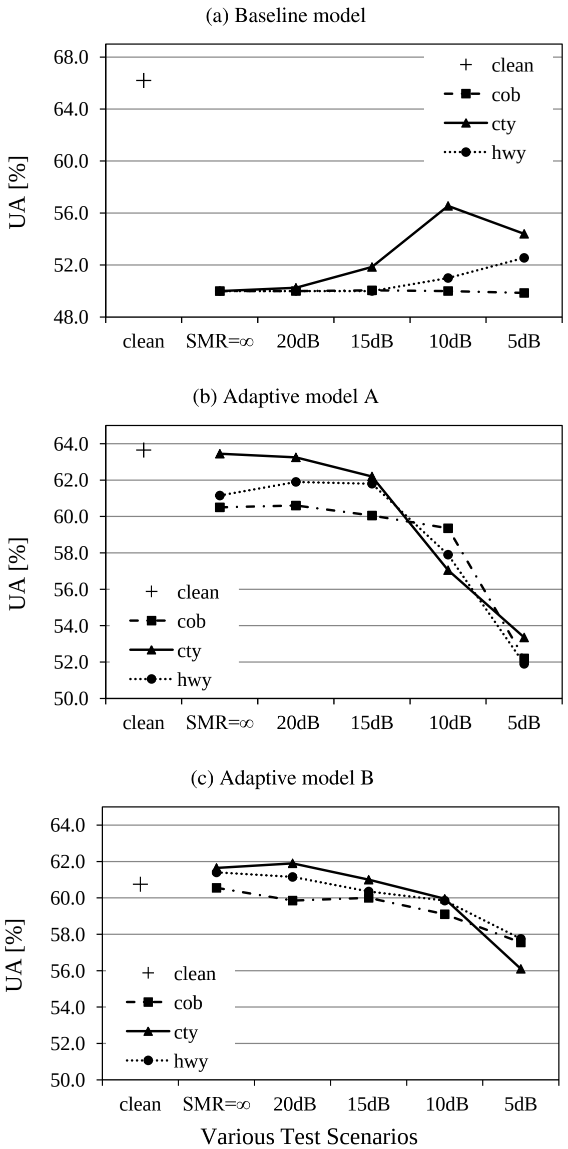
3.3 Acoustic Model Training

In accordance with the 2011 SSC evaluation procotol, results are evaluated in terms of unweighted accuracy (UA), which is simply the average of the recalls of the 'AL' and 'NAL' classes. Corresponding to the 2011 SSC baseline, our first acoustic model is only trained on the clean training and development sets (9 360 utterances) without any distortion by any type of intrinsic in-car noise and background music. Figure 2 (a) depicts the unweighted accuracies (UAs) with this baseline model when testing on 16 test sets with different acoustic situations (cf. Subsection 3.2). The best performance (66.2 % UA) can be observed in case of testing on clean data only. However, from Figure 2 (a), one can see that the model classifies with almost chance level accuracy (near 50 % UA) when testing on sets containing any type of noise.

In order to adapt to various in-car acoustic environments, multi-condition training is considered to improve the robustness of models. In our experiments we employ the following two multi-condition training strategies:

[1] Thanks to Martin Wöllmer for providing the driving noise corpus.

Figure 2: *Unweighted accuracy (UA) for binary intoxication states (AL / NAL) classification training on 3 acoustic models: (a) Baseline model; (b) Adaptive model A: training with in-car driving noise adaptation; (c) Adaptive model B: training with in-car driving noise and background music. SMR: speech-to-music ratio. Testing in 16 conditions: clean; with driving noise only (cobble, city, highway at SMR=∞); driving noise and music (SMR < ∞).*



1) **Adaptive model A**: training on clean (scenario 1) and in-car driving noise data (scenario 2);

2) **Adaptive model B**: training on clean (scenario 1), in-car driving noise (scenario 2), and additive background music data (scenario 3).

It is ensured that all training sets have equal size by random down-sampling of the multi-condition training sets. More precisely, the training set for building adaptive model A is defined by randomly selecting 9 360 utterances equally and without overlap from the train and development partitions of the four corpora from scenario 1 and scenarios 2. Consequently, 2 340 utterances are selected from each condition (clean, CTY, HWY, and COB noise). Analogously, the training set for building adaptive model B is defined by random selection of 9 360 utterances from the train and development partitions of all 16 corpora from scenarios 1–3. Hence, 585 utterances are selected from each corpus, so that the SNR levels in training are balanced. Therefore, the acoustic model A aims to adapt to in-car noise, and acoustic model B tries to eliminate the influences of both intrinsic noise and music. Instead of training and testing on 'matched condition' with single type of in-car noise, the multi-condition trained models are more connected to a realistic application where the acoustic conditions change frequently.

Figures 2 (b) and (c) display the performance for intoxication state classification based on the above mentioned two adaptive acoustic models (A and B). As is evident from Figure 2 (b), most of the UAs are significantly improved when testing on any kind of noise, as compared to Figure 2 (a). For example, the UAs rise up to 63.5 %, 61.2 % and 60.5 % from 50.0 % when testing on data only overlaid by one of the CTY, HWY, and COB noise types, respectively. This can be explained by the fact that adaptive model A includes more distortion information caused by driving—as the training set is of equal size as for the baseline, it cannot be simply attributed to more training data. However, the performance degrades seriously as soon as music is added at lower SMRs. To alleviate this problem, employing adaptive model B seems reasonable. From Figure 2 (c), it can be seen that the performance is significantly[2] enhanced by up to six percent absolute (from 53.4 %, 51.9 %, and 52.2 % to 56.1 %, 57.8 % and 57.6 % UA for CTY, HWY and COB) in case of testing on data at an SMR of 5 dB.

However, from Figure 2 (b) and (c), one can notice that the UAs for testing on the clean set are significantly lower than the baseline result (from 66.2 % to 63.7 % for adaptive model A; to 60.8 % for adaptive model B). Therefore, we can conclude that the enhanced performance of the multi-condition models in noisy conditions comes at the price of degradations on clean data.

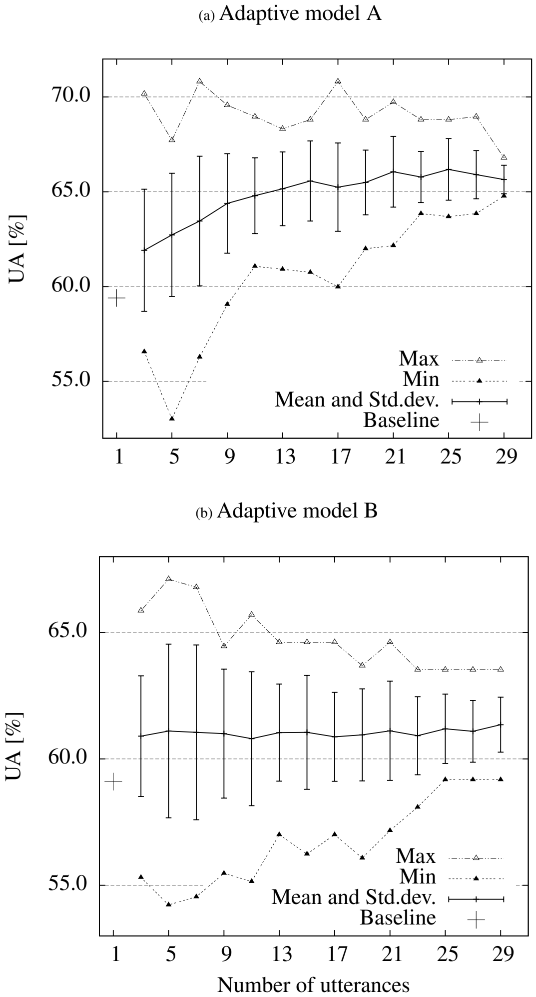
Still, considering the average performance across all test conditions (which is roughly the 'expected performance' in realistic conditions), model A and model B significantly outperform baseline model without any adaptation, as their average UAs across the 16 test scenarios are 59.9 % and 59.4 % in comparison to 52.0 %. Furthermore, it can be seen that while the performance on model B is more stable across SNRs than model A, its overall performance is lower. A general trend is that the COB noise type affects the performance most seriously, followed by HWY noise and CTY noise, which corresponds to their SNRs (cf. Figure 1) and matches previous studies on ASR in these conditions [9].

3.4 Fusing Utterance Level Decisions

As discussed above, model 'adaptation' for multiple acoustic environments is somewhat effective to overcome the impact of in-car noise on intoxication state detection. It can be argued, however, that the performance is still not sufficient for realistic applications. Thus, in the following, we continue to investigate the strategy of fusing utterance level decisions along the time axis, as presented in [8], to profit from temporal 'smoothing' of classifier decisions in a 'session' where the speaker BAC is assumed to be constant.

[2]As a rule of thumb for the ALC test set, improvements/degradations of above 2.5 % absolute are significant at the 5 % level according to a one-tailed z-test.

Figure 3: *Unweighted accuracy (UA) for binary intoxication (AL / NAL) classification when testing on speech set overlaid by COB noise at SMR = 10 dB vs. number of fused utterance level decisions (3–29) from two adaptive acoustic models (cf. Subsection 3.3).*



As done in [8], for each speaker the unweighted majority vote is taken over N randomly selected utterances from each of the alcoholized and non-alcoholized sessions[3] to determine a decision for each session; then, the session level UA is computed. The parameter N is chosen as an odd number from $\{3, \ldots, 29\}$ to suit for majority voting. Further, 30 iterations of the experiments are performed for reducing the effect of statistical fluctuations.

In the following, we take the test case of SMR = 10 dB with COB noise type out of the 16 possible cases as an example. Figure 3 shows the UA distribution for majority votes over an increasing number of utterance level decisions. Both adaptive model A (Fig. 3 (a)) and B (Fig. 3 (b)) are employed to alleviate the noise influence, as discussed in the above subsection. The baseline mean UAs of 59.9 % / 59.4 % for adaptive model A / B are roughly equal to the expected UAs measured on session level when randomly picking a single utterance per session.

For both models, the performance can be increased by 2.5 % and 1.7 % absolute by fusing three utterances. Further, for model A, the expected mean UA can be constantly improved by accumulating utterance level decisions. The best performance is achieved at 66.2 % mean UA, $N = 25$. The trend line is similar to the result in the clean case [8]. However, for model B, the improvement is not obvious for $N > 3$. By testing all other possible sets (not displayed), we find that for increased noise levels, the improvement by the majority voting strategy will be less and less. In turn, this probably indicates that the predictions of multi-condition trained models are less consistent.

[3] Note that random selection is used since the ALC utterances are spoken out-of-context.

4 Conclusions and Future Work

We have investigated the influence of driving noise and background music on the accuracy of automatic alcohol intoxication recognition from speech. In a large scale study we have demonstrated that the accuracy severely degrades when we use the baseline acoustic model trained on clean data in mismatched conditions. To enhance the robustness of possible vehicle-mounted systems and mitigate performance degradation, we investigated multi-condition training and majority voting along the time axis. Such fusion corresponds to possible in-car applications with long-term observation of the driver rather than, e. g., a single alcohol test when starting the car. In the result, 66.2 % UA can be expected by combining both strategies in adverse acoustic conditions (SMR of 10 dB and additional noise from a cobble road). Future work should focus on noise-robust and context-sensitive recognition architectures such as recurrent neural networks.

References

[1] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[2] M. Sakairi and M. Togami, "Use of water cluster detector for preventing drunk and drowsy driving," in *Proc. Sensors, IEEE*, (Waikoloa, HI, USA), pp. 141–144, November 2010.

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

[4] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. INTERSPEECH 2011*, (Florence), pp. 3201–3204, ISCA, 2011.

[5] U. Remes, K. J. Palomäki, and M. Kurimo, "Robust automatic speech recognition using acoustic model adaptation prior to missing feature reconstruction," in *Proc. of the 17th European Signal Processing Conference, EUSIPCO 2009*, (Glasgow, UK), pp. 535–539, 2009.

[6] K. Yao, K. K. Paliwal, and S. Nakamura, "Noise adaptive speech recognition based on sequential noise parameter estimation," *Speech Communication*, vol. 42, no. 1, pp. 5–23, 2004.

[7] F. Weninger, M. Wöllmer, and B. Schuller, "Combining Bottleneck-BLSTM and semi-supervised sparse NMF for recognition of conversational speech in highly instationary noise," in *Proc. INTERSPEECH 2012*, (Portland, OR, USA), ISCA, 2012. To appear.

[8] F. Weninger, E. Marchi, and B. Schuller, "Improving Recognition of Speaker States and Traits by Cumulative Evidence: Intoxication, Sleepiness, Age and Gender," in *Proc. INTERSPEECH 2012*, (Portland, OR, USA), 2012, to appear.

[9] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, and G. Rigoll, "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," in *Proc. of ICASSP*, (Dallas, TX, USA), pp. 4562–4565, 2010.