Detecting Road Surface Wetness from Audio: A Deep Learning Approach

Irman Abdić, Lex Fridman, Daniel E. Brown, William Angell, Bryan Reimer
Massachusetts Institute of Technology
Cambridge, MA, USA
abdic,fridman,danbr,wha,reimer@mit.edu

Erik Marchi
Technische Universität München
Munich, Germany
erik.marchi@tum.de

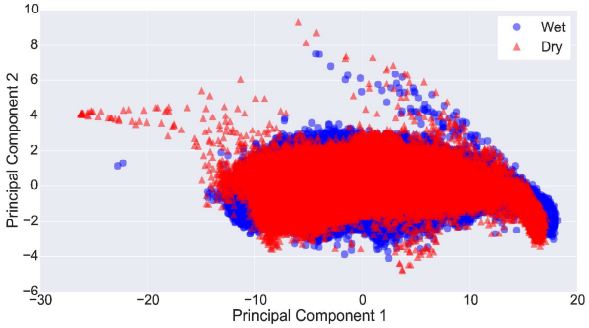
Björn Schuller
Imperial College London
London, UK
schuller@ieee.org

Abstract

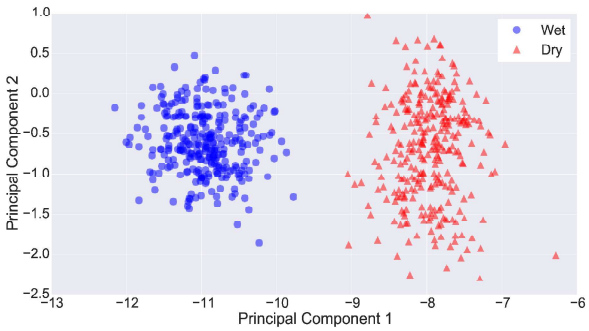
We introduce a recurrent neural network architecture for automated road surface wetness detection from audio of tire-surface interaction. The robustness of our approach is evaluated on 785,826 bins of audio that span an extensive range of vehicle speeds, noises from the environment, road surface types, and pavement conditions including international roughness index (IRI) values from 25 in/mi to 1400 in/mi. The training and evaluation of the model are performed on different roads to minimize the impact of environmental and other external factors on the accuracy of the classification. We achieve an unweighted average recall (UAR) of 93.2 % across all vehicle speeds including 0 mph. The classifier still works at 0 mph because the discriminating signal is present in the sound of other vehicles driving by.

1. Introduction

Wet pavement is responsible for 74 % of all weather-related crashes in the U.S. with over 380,000 injuries and 4,700 deaths per year [7]. Furthermore, wet roads often increase traffic congestion and result in infrastructure damage and supply chain disruptions [3]. From the perspective of driver safety, wetness detection during the period of time after the percipitation has ended but whether the road is still wet is critical. Under these conditions, human estimation of road wetness and friction properties is less accurate than normal, especially in reduced visibility over night or in the presence of fog [4].



(a) PCA for selected full wet and dry trips from Table 1.



(b) PCA for a randomly-selected segment of road from wet and dry trips from Table 1.

Figure 1: PCA analysis for wet and dry road surfaces that illustrates a representative case where audio-based wetness detection is linearly separable for similar road type and vehicle speeds.

The automated detection of road conditions from au-

dio may be an important component of next generation Advanced Driver Assistance Systems (ADAS) that have the potential to enhance driver safety [26]. Moreover, autonomous and semi-autonomous vehicles have to be aware of road conditions to automatically adapt vehicle speed while entering the curve or keep a safe distance to the vehicle in front. There are numerous approaches that can detect whether a surface is wet or dry, but in the majority of cases they are not robust to variation in real-world datasets. Accuracy of video-based wetness prediction decreases significantly in poor lighting conditions (i.e., night, fog, smoke). Audio-based wetness prediction is heavily dependent upon surface type and vehicle speed which is fairly represented in our dataset of 785,826 bins (feature vectors described in §2.2). We elucidate this dependence by visualizing the first two principal components for (1) two full trips and (2) a small 10-second section of road from (1). These two visualizations are shown in Fig. 1a and Fig. 1b, respectively. The feature set we use is linearly separable for a specific road type and vehicle speed, as visualized in Fig. 1b. However, given the nonlinear relation of our feature set for (1) that is visualized in Fig. 1a we applied Recurrent Neural Networks (RNNs) which can model and separate the data points.

1.1 Related Work

Long short-term memory RNNs (LSTM-RNNs) have been successfully applied in many fields from hand writing recognition to robotic heart surgery [11, 25]. In the audio context, LSTM-RNNs contributed to the development of better phoneme classification, speech enhancement, affect recognition from speech, animal species identification and finding temporal structure in music [8,12,22,29,31,32]. However, to our best knowledge LSTM-RNNs have not been applied to the task of road wetness detection.

Related works can be found in the video processing domain, where wetness detection has been studied with two camera set-ups: (1) a surveillance camera at night, and (2) a camera on-board a vehicle. The detection of road surface wetness using surveillance camera images at night is relying on passing cars' headlights as a lighting source that creates a reflection artifact on the road area [17]. On-board video cameras use polarization changes of reflections on road surfaces or spatio-temporal reflection models [2, 19, 33]. A recent study uses near infrared (NIR) camera to classify several road conditions per every pixel with a high accuracy, the evaluation has been done in laboratory conditions, and field experiments [20]. However, a drawback of video processing methods is that they require (1) an external illumination source to be present and (2) visibility conditions to be clear.

Another approach capable of detecting road wetness relies on 24-GHz automotive radar technology for detecting low-friction spots [28]. It analyzes backscattering properties of wet, dry, and icy asphalt in laboratory and field experiments.

Traditionally, audio analysis of the road-tire interaction has been done by examining tire noises of passing vehicles from a stationary microphone positioned on the side of the road. This kind of analysis reveals that tire speed, vertical tire load, inflation pressure and driving torque are primary contributors to tire sound in dry road conditions [18]. Acoustic-based vehicle detection methods, as the one that uses bispectral entropy have been applied in the ground surveillance systems [5]. Other on-road audio collecting devices for surface analysis can be found in specialized vehicles for pavement quality evaluation (e.g., VOTERS [6]) and for vehicles instrumented for studying driver behavior in the context of automation (e.g., MIT RIDER [10]). Finally, road wetness has been studied from on-board audio of tire-surface interaction, where SVMs have been applied [1].

1.2 Contribution

The method described in our paper improves the prediction accuracy of the method presented in [1] and expands the evaluation to a wider range of surface types and pavement conditions. Additionally, the present study is the first in applying context aware LSTM-RNNs in this field. Moreover, we improve on the following three aspects of [1] where (1) the model was trained and tested on the same road segment, (2) false predictions caused by the impact of pebbles on the vehicle chassis were ignored, and (3) audio segments associated with speeds below 18.6 mph were removed.

We trained and tested the model on different routes, and considered all predictions regardless of the speed, pebbles impact or any other factor.

2 Road Surface Wetness Classification

2.1 Data Collection

For data collection purposes, we instrumented a 2014 Mercedes CLA with an inexpensive shotgun microphone behind the rear tire, as shown in Fig. 2. The gain level of the microphone and its distance from the tire were kept the same for the entire data collection process. Three different routes were selected. For each route, we drove the same exact path once during the rain (or immediately after) and another time when the

road surface was completely dry, as shown in Fig. 3. We provide spectrograms in Fig. 4 for wet and dry road segments of the same route that highlight the difference in frequency response. The duration and length of trips ranged from 14 min to 30 min and 6.1 mi to 9.0 mi, respectively. The summary of the dataset is presented in Table 1.

Figure 2: Instrumented MIT AgeLab vehicle (left) and placement of the shotgun microphone behind the rear tire (right).

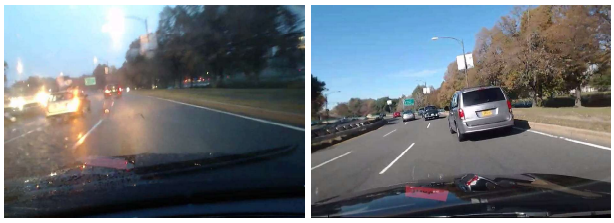


Figure 3: Snapshots from the video of the forward roadway showing the same GPS location for a 'wet' trip 1 (left) and a 'dry' trip 1 (right). [a]

[a] A video of these trips is available at: http://lexfridman.com/wetroad

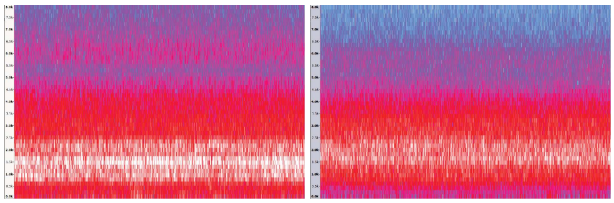


Figure 4: Spectrograms for the wet trip 2 (left) and dry trip 2 (right) from the same route segment at the speed of approximately 20 mph.

Trip	Time	Distance	Avg Speed	Avg IRI
wet 1	26 min	9.0 mi	7.4 mph	267 in/mi
wet 2	16 min	6.4 mi	9.4 mph	189 in/mi
wet 3	14 min	6.1 mi	13.5 mph	142 in/mi
dry 1	30 min	9.0 mi	9.6 mph	267 in/mi
dry 2	14 min	6.4 mi	9.1 mph	189 in/mi
dry 3	18 min	6.1 mi	9.3 mph	142 in/mi

Table 1: Statistics of the collected data for six trips: time, distance, average speed and average IRI.

The data collection was carried out in Cambridge and the Greater Boston area with different speeds, traffic conditions and pavement roughness. The latter is measured with the International Roughness Index (IRI) which represents pavement quality [27]. A histogram of IRI values for the collected dataset is presented in Fig. 5, wherein the unit of measurement is in inches per mile (in/mi). Our dataset contains values from 25 in/mi to 1400 in/mi, but in Fig. 5, values over 400 in/mi are aggregated into a single bin. According to the Massachusetts Department of Transportation (MassDOT) Road Inventory, the route we traveled is a combination of surface-treated road and bituminous concrete road [24].

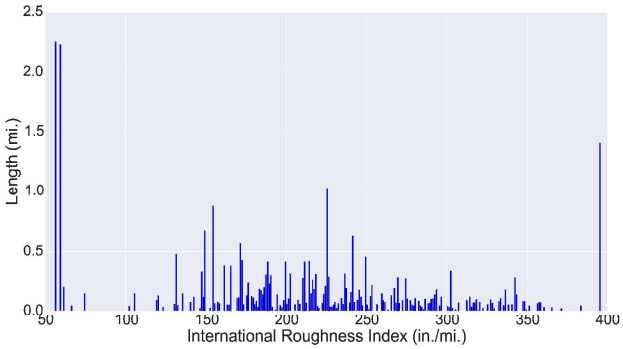


Figure 5: Histogram or IRI distribution throughout collected data.

2.2 Features

Our aim was to model the whole spectrum along with the first order differences and then select a subset of features that discriminates our classes the best. In contrast with the previous work from [1] which uses 125 ms frame size, we extracted Auditory Spectral Features (ASF) [22], that were computed by applying the short-time Fourier transform (STFT) using a frame size 30 ms to increase the precision in describing the data, and a frame step of 10 ms to produce more observations. Furthermore, each STFT power spectrogram has been converted to the Mel-Frequency scale using 26 triangular filters obtaining the Mel spectrograms $M^{30}(n,m)$. As humans are able to recognize if road is wet or dry from audio, we chose a logarithmic representation to match the human perception of loudness:

$$M^{30}_{log}(n,m) = log(M^{30}(n,m) + 1.0), \qquad (1)$$

where n is a frame index, and m is a frequency bin index. In addition, the positive first order differences

$D^{30}(n,m)$ were calculated from each Mel spectrogram as follows:

$$D^{30}(n,m) = M^{30}_{log}(n,m) - M^{30}_{log}(n-1,m). \quad (2)$$

The frame energy has also been included as a feature which resulted in a total of 54 features [23]. We started with the great representation of the spectrum by using all 54 features to provide a good modelling and then we ranked the features to reduce the dimensionality. To foster reproducibility, we use the opensource software toolkits: (a) openSMILE – for extracting features from the audio, and (b) Weka 3 – for feature evaluation with Information Gain (IG) and Correlation-based Feature Selection (CFS) to reduce the dimension of the feature space [9, 13].

The IG feature evaluation is an univariate filter that calculates the worth of a feature by measuring the IG with respect to the class, it measures individual feature value but neglects redundancy [14, 21]. The output is a list of ranked features of which we selected best $5n$ features, where $n \in [1..10]$ and the whole feature set for comparison.

The CFS subset evaluation is a multivariate filter that seeks for subsets of features that are highly correlated with the class while having low intercorrelation [13, 15, 21]. We used the BestFirst search algorithm in a forward search mode (-D 1) and a threshold of 5 non-improving nodes (-N 5) for consideration before terminating search. The CFS subset evaluation was applied on ASF considering all bins and it returned a list of 5 features.

2.3 Classifier

In this work, we used a deep learning approach with initialized nets – LSTM and bi-directional LSTM (BLSTM) RNN architectures which in contrast to other RNNs do not suffer from the problem of vanishing gradients [16]. The BLSTM is an extension of the LSTM architecture that allows for an additional forward pass if a look-ahead buffer may be used, which has been proven successful in many applications [12].

In addition, we evaluated different parameters, such as the layout of LSTM and BLSTM hidden layers (54-54-54, 54-30-54, 156-256-156, 216-216-216, 216-316-216 neurons in the three hidden layers) and learning rates (1e-4, 1e-5, 1e-6). Initially, we chose deep architecture with three hidden layers of the same size as input vectors (54), before we ranked features and reduced its dimensionality. In the next step we investigated effectiveness of internal feature compression and augmentation of hidden layers to model more information. We used feed forward output layer with a logistic activation function and sum of squared error as objective function. The experiments were carried out with the CURRENNT toolkit [30].

3 Results

Table 2 shows the evaluation results in an ascending order for the best 20 features that were selected with IG (IG-20), as described in §2.2 and trained with LSTM-RNNs. We present only the worst three and the best three results for LSTM-RNNs, whereas other experiments were left out from the table. For every combination of parameters we conducted cross-validation on all three folds from Table 1. I.e., we leave out wet/dry 3 at a time for training with wet/dry 1 and testing with wet/dry 2, and run six experiments in total. Furthermore, an average UAR was computed for results obtained from all speeds including vehicle stationary mode. The best result with an UAR of 93.2 % was achieved with BLSTM network layout 216-216-216 and learning rate $1e^{-5}$.

Additionally, we compared our results with the state-of-the-art approach of [1] that uses zero-norm minimization (L0) to select four most promising features (L0-4) from 125 ms audio bins of 1/3 octave bands (5000 Hz, 1600 Hz, 630 Hz and 200 Hz frequency bands). We trained SVMs with Sequential Minimal Optimization (SMO) on our dataset and found a C parameter of $1e^{-3}$ to give the best UAR of 67.4 %. Furthermore, experiments with SVMs and IG-20 feature set were carried out and gave the best UAR of 78.8 %.

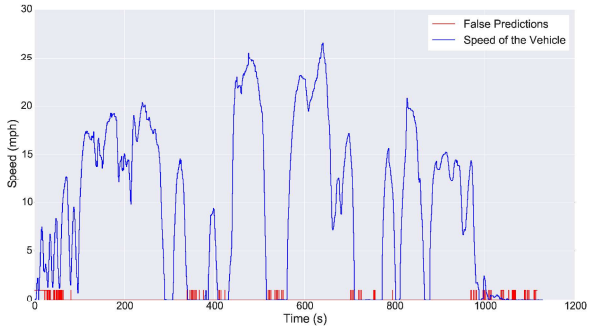
Network	Feature set	C ($1e^{-n}$)	UAR (%)
SVM	Z0-4	3	67.4
SVM	IG-20	3	78.8
Network	**Layout**	**LR ($1e^{-n}$)**	**UAR (%)**
LSTM	216-216-216	4	66.3
BLSTM	216-316-216	4	76.1
LSTM	156-256-156	4	78.0
⋮	⋮	⋮	⋮
BLSTM	216-316-216	5	92.6
LSTM	216-216-216	5	92.6
BLSTM	**216-216-216**	**5**	**93.2**

Table 2: Comparison of results (upper) that were obtained by applying state-of-the-art approach of Alonso *et al.*, and (lower) our approach with LSTM-RNNs, both trained and tested on our dataset. The column LR is an abbreviation for Learning Rate.

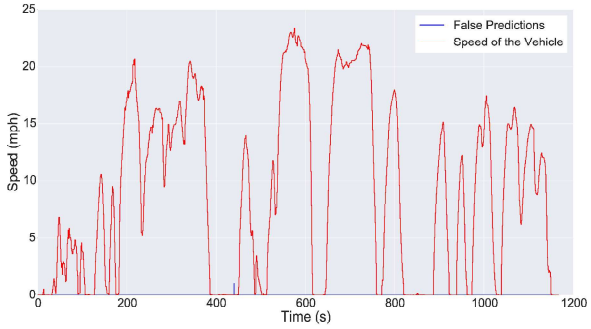
The mean UAR value for experiments with LSTM-RNNs is 86.6 % and the standard deviation equals 6.4. The mean UAR of all experiments with BLSTM network is 87.0 %, while the mean UAR for experiments

with LSTM network is 86.0 %. The best mean UAR for experiments with learning rate $1e^{-5}$ amounts to 90.8 %, while the worst performing learning rate $1e^{-4}$ achieves only 78.8 %.

Two out of three wet trips have significantly higher number of false predictions (1) at the beginning, where vehicle tires were dry before getting wetted from the surface, and (2) at the end of the trip, when the vehicle entered a parking lot with relatively dry road surface.



(a) An 18 min long wet trip showing speed and false predictions.



(b) A 19 min long dry trip showing speed and false predictions.

Figure 6: Graphs for wet and dry road surfaces for route two that clarify the correlation between low speed and inaccurate predictions.

In Fig. 6 we compare speed and false predictions of wet and dry trips for the same route that has similar properties, which are described in §2.1. One can observe that all false predictions of wet trip 2 in Fig. 6a occured below the speed of 2.9 mph, whilst Fig. 6b depicts a dry trip 2 and has only one false prediction when the vehicle is not moving. Therefore, discarding speeds below 2.9 mph improves the UAR to 100 %. When we look only at speeds below 2.9 mph and ignore everything above we are still able to attain 74.5 % UAR. The latter is possible only in presence of ambient sounds, as noises of vehicles that are driving by.

4 Conclusion

We proposed a deep learning approach based on LSTM-RNNs for detecting road wetness from audio of the tire-surface interaction and discriminating between wet and dry classes. This method is shown to be robust to vehicle speed, road type, and pavement quality on a dataset containing 785,826 bins of audio. It outperforms the state-of-the-art SVMs and achieves an outstanding performance on the road wetness detection task with an 93.2 % UAR for all vehicle speeds and the more challenging speeds being those below 2.9 mph, including vehicle stationary mode. In future work, we will increase the variability in the audio observations by collecting the data with different vehicles, tire models, tire wears, and inflation pressure. Additionally, we will augment the feature set for estimating depth of water on the road surface and detecting hydroplaning conditions.

Acknowledgments

Support for this work was provided by the New England University Transportation Center, and the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs class counsel.

References

[1] J. Alonso, J. López, I. Pavón, M. Recuero, C. Asensio, G. Arcas, and A. Bravo. On-board wet road surface identification using tyre/road noise and support vector machines. *Applied Acoustics*, 76:407–415, 2014.

[2] M. Amthor, B. Hartmann, and J. Denzler. Road condition estimation based on spatio-temporal reflection models, 2015.

[3] J. Andrey, B. Mills, M. Leahy, and J. Suggett. Weather as a chronic hazard for road transportation in canadian cities. *Natural Hazards*, 28(2-3):319–343, 2003.

[4] J. Andrey, B. Mills, and J. Vandermolen. Weather information and road safety. *Institute for Catastrophic Loss Reduction, Toronto, Ontario, Canada*, 2001.

[5] M. Bao, C. Zheng, X. Li, J. Yang, and J. Tian. Acoustical vehicle detection based on bispectral entropy. *Signal Processing Letters, IEEE*, 16(5):378–381, 2009.

[6] R. Birken, G. Schirner, and M. Wang. Voters: design of a mobile multi-modal multi-sensor system. In *Proceedings of the Sixth International Workshop on Knowledge Discovery from Sensor Data*, pages 8–15. ACM, 2012.

[7] Booz-Allen-Hamilton. Ten-year averages from 2002 to 2012 based on nhtsa data. *US Department of Transportation - Federal Highway Administration*, 2012.

[8] D. Eck and J. Schmidhuber. Finding temporal structure in music: Blues improvisation with lstm recurrent networks. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 747–756. IEEE, 2002.

[9] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

[10] L. Fridman, D. E. Brown, W. Angell, I. Abdić, B. Reimer, and H. Y. Noh. Automated synchronization of driving data using vibration and steering events. *Pattern Recognition Letters*, 2016, In Press.

[11] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):855–868, 2009.

[12] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[14] M. Hall, G. Holmes, et al. Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 15(6):1437–1447, 2003.

[15] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] Y. Horita, S. Kawai, T. Furukane, and K. Shibata. Efficient distinction of road surface conditions using surveillance camera images in night time. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 485–488. IEEE, 2012.

[18] K. Iwao and I. Yamazaki. A study on the mechanism of tire/road noise. *JSAE review*, 17(2):139–144, 1996.

[19] M. Jokela, M. Kutila, and L. Le. Road condition monitoring system based on a stereo camera. In *Intelligent Computer Communication and Processing, 2009. ICCP 2009. IEEE 5th International Conference on*, pages 423–428. IEEE, 2009.

[20] P. Jonsson, J. Casselgren, and B. Thornberg. Road surface status classification using spectral analysis of nir camera images. *Sensors Journal, IEEE*, 15(3):1641–1656, 2015.

[21] A. G. Karegowda, A. Manjunath, and M. Jayaram. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277, 2010.

[22] E. Marchi, G. Ferroni, F. Eyben, L. Gabrielli, S. Squartini, and B. Schuller. Multi-resolution Linear Prediction Based Features for Audio Onset Detection with Bidirectional LSTM Neural Networks. In *Proceedings 39th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014*, pages 2183–2187, Florence, Italy, May 2014. IEEE, IEEE. (acceptance rate: 50 %, IF* 1.16 (2010)).

[23] E. Marchi, F. Vesperini, F. Weninger, F. Eyben, S. Squartini, and B. Schuller. Non-Linear Prediction with LSTM Recurrent Neural Networks for Acoustic Novelty Detection. In *Proceedings 2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, Killarney, Ireland, July 2015. IEEE, IEEE.

[24] MassDOT. Road inventory - massdot planning, 2015.

[25] H. Mayer, F. Gomez, D. Wierstra, I. Nagy, A. Knoll, and J. Schmidhuber. A system for robotic heart surgery that learns to tie knots using recurrent neural networks. *Advanced Robotics*, 22(13-14):1521–1537, 2008.

[26] M. Mueller. Sensor sensibility: Advanced driver assistance systems. *Vision Zero International*, 2015.

[27] W. D. Paterson. International roughness index: Relationship to other measures of roughness and riding quality. *Transportation Research Record*, (1084), 1986.

[28] V. V. Viikari, T. Varpula, and M. Kantanen. Road-condition recognition using 24-ghz automotive radar. *Intelligent Transportation Systems, IEEE Transactions on*, 10(4):639–648, 2009.

[29] F. Weninger and B. Schuller. Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations. In *acoustics, speech and signal processing (ICASSP), 2011 IEEE international conference on*, pages 337–340. IEEE, 2011.

[30] J. Weninger, Felix Bergmann and B. Schuller. Introducing currennt - the munich open-source cuda recurrent neural network toolkit. *Journal of Machine Learning Research*, (16):547–551, 2014.

[31] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, volume 2008, pages 597–600, 2008.

[32] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee. An experimental study on speech enhancement based on deep neural networks. *Signal Processing Letters, IEEE*, 21(1):65–68, 2014.

[33] M. Yamada, T. Oshima, K. Ueda, I. Horiba, and S. Yamamoto. A study of the road surface condition detection technique for deployment on a vehicle. *JSAE review*, 24(2):183–188, 2003.