Acoustic Feature Analysis for Wet and Dry Road Surface Classification Using Two-stream CNN

Siavash Bahrami
Universiti Putra Malaysia
siavash.bahrami@live.com

Shyamala Doraisamy
Universiti Putra Malaysia
shyamala@upm.edu.my

Azreen Azman
Universiti Putra Malaysia
azreenazman@upm.edu.my

Nurul Amelina Nasharuddin
Universiti Putra Malaysia
nurulamelina@upm.edu.my

Shigang Yue
University of Lincoln
syue@lincoln.ac.uk

ABSTRACT

Road surface wetness affects road safety and is one of the main reasons for weather-related accidents. Study on road surface classification is not only vital for future driverless vehicles but also important to the development of current vehicle active safety systems. In recent years, studies on road surface wetness classification using acoustic signals have been on the rise. Detection of road surface wetness from acoustic signals involve analysis of signal changes over time and frequency-domain caused by interaction of the tyre and the wet road surface to determine the suitable features. In this paper, two single stream CNN architectures have been investigated. The first architecture uses MFCCs and the other uses temporal and spectral features as the input for road surface wetness detection. A two-stream CNN architecture that merges the MFCCs and spectral feature sets by concatenating the outputs of the two streams is proposed for further improving classification performance of road surface wetness detection. Acoustic signals of wet and dry road surface conditions were recorded with two microphones instrumented on two different cars in a controlled environment. Experimentation and comparative performance evaluations against single stream architectures and the two-stream architecture were performed. Results shows that the accuracy performance of the proposed two-stream CNN architecture is significantly higher compared to single stream CNN for road surface wetness detection.

CCS CONCEPTS

• **Computing methodologies** → Machine learning; Machine learning algorithms; Feature selection.

KEYWORDS

Acoustic signal processing, Artificial neural networks, Feature selection, Intelligent transportation systems, Vehicle safety

CSAI 2020, December 11–13, 2020, Zhuhai, China

ACM ISBN 978-1-4503-8843-6/20/12…$15.00
https://doi.org/10.1145/3445815.3445847

ACM Reference Format:
Siavash Bahrami, Shyamala Doraisamy, Azreen Azman, Nurul Amelina Nasharuddin, and Shigang Yue. 2020. Acoustic Feature Analysis for Wet and Dry Road Surface Classification Using Two-stream CNN. In *2020 4th International Conference on Computer Science and Artificial Intelligence (CSAI 2020), December 11–13, 2020, Zhuhai, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3445815.3445847

1 INTRODUCTION

Slippery wet road surfaces is one of the major causes of weather-related accidents. Reports shows that in 2016, 25,777 weather related fatal accidents had happened in Europe [1]. Vehicle stability on the road depends on the friction forces of the tyre to the road surface and this is considered as an important factor towards the development of vehicle safety systems [2]. Knowledge of road surface wetness condition is an important factor in designing an active safety feature for autonomous and semi-autonomous vehicles. Using this knowledge, the vehicle can automatically adjust its speed to maintain a safe distance to the front vehicle and also have better manoeuvrability on slippery roads enabling better driver safety assistant systems [3].

Experiments using mounted microphones and tyre sound recordings have been on the rise for detecting road surface wetness. Using acoustic signals to determine road surface conditions was first successfully started by installing the microphone on the side of the road [4] and collecting sound recordings from passing vehicles. These recording samples are not a complete representation of road condition as only a few sections of the road are recorded. More comprehensive recording samples can be found in recent studies where data collection is done on-board of the vehicle [5]–[9]. Using acoustic signals is also useful to overcome the poor illumination where computer vision system performance might be affected. This study can be integrated alongside systems such as [10], [11] where one of the challenges with these systems is the need for external illumination and may perform poorly in low light conditions.

Various machine learning algorithms have been used for road wetness classification, such as studies by [5] and [8] using support vector machines (SVM) and recurrent neural networks (RNN) Long short-term memory (LSTM) and Bidirectional-LSTM(BLSTM) used by [7]. CNN was studied by [9] and artificial neural networks (ANN) were investigated by [6].

CNNs promising classification performance has also been demonstrated in other acoustic signal classification tasks such as environment sound classification [12]–[15]. [13], [15] use a stacked

CNN model for environmental event sound recognition, where mel-spectrogram and raw audio features are fed into two separate CNNs and then combined using Dempster-Shafer (DS) fusion algorithm to get the final prediction. [12] Trained a 4-layers CNN model consisting of 2 convolution layers with max-pooling and 2 dense layers. The feature set used in [12] was deltas of segmented spectrograms. The Network proposed by [14] comprised three convolutional layers interleaved with two pooling layers, followed by two fully connected layers with 64 and 10 hidden unit respectively. In addition, [14] investigated experiments with data augmentation. Based on the high classification performance of the CNNs on sound classification, in this paper, we investigate the use of two-stream CNN architectures for road wetness detection to improve classification performance.

Most of the studies on road surface wetness detection had used octave-band features [5], [8] or auditory spectral features (ASF) [7], [9] from acoustic signals.

Research on environment sound classification using multi-stream CNN architecture [13], [15] has been investigated. We propose a two-stream CNN architecture for road wetness classification that consist of two separate 4-layer CNNs were trained using different feature sets as the input. One CNN stream uses 13 MFCCs as the input and we refer to it as MFCCs-CNN and the other stream uses time domain and frequency domain features as the input and we refer to it as Spectral-CNN. The extracted feature map resulting from each stream of the two-stream CNN are then concatenated and fed to fully connected layers for classification.

The rest of the paper is structured as follows. Section 2 presents a general background on feature extraction and Section 3 is a description of the proposed architectures. Section 4 presents the experimental setup and results. Finally, conclusions and future research directions are given in Section 5.

2 BACKGROUND

In this section, a background discussion on acoustic signal features and recent works and techniques that can be used for feature learning is given.

2.1 Acoustic Features

Detection of road surface wetness from acoustic signals involves the analysis of the signal changes over time and frequency-domain caused by interaction of tyre and the road surface to determine the features that can be used. The acoustic signals classification performance of all machine learning approaches depends on the extraction and selection of features related to the specific task. Most of the acoustic features are designed for tasks such as speech or music. Finding features that can be used to effectively differentiate wet and dry road surface acoustic signals is a challenging task. Acoustic features in general can be represented in time, frequency, or time-frequency domains. In this paper, the spectral and temporal features from each domain have been selected for road surface wetness classification, based on recent research carried on environment sound recognition task [12]–[14], [16].

In the time domain, the features selected are root mean square energy (RMSE) and zero crossing rate (ZCR). As for the frequency domain audio spectrum flatness (ASF), spectral contrast, spectral centroid and spectral roll-off have been selected. In this paper, we refer to this feature set as RZASSR.

The mel-frequency cepstrum coefficients (MFCCs) is the most popular cepstrum-based audio features in time-frequency domains. MFCC can be defined as a short-window cepstrum of a signal and is used in different acoustic classification tasks such as speech, music and environment sound. A detailed description of these features can be found in [16].

Figure 1 shows an example of the RMSEs and spectral centroids visualized in time domain, and visualization of MFCCs in time-frequency domain extracted from 2 minutes of acoustic signals. These acoustic signals are part of our dataset, recorded from wet and dry road surfaces using mounted microphones on a car while driving on the same road.

The description of this dataset will be discussed in Section 4. The visualizations of the RMSE and spectral centroid in Figure 1 (a) and MFCCs in Figure 1 (b) show a clear difference between wet and dry road surface acoustic signals and it requires further analysis. Next, we will discuss feature learning and how these features can be used to classify wet and dry road surfaces.

2.2 Feature Learning With CNN

Feature learning techniques allow a machine learning model to discover the features needed for classification from raw data automatically. Feature learning has shown promising results in acoustic scene classification tasks. Road surface acoustic classification task has a similar characteristic to the acoustic scene where a wide variety of possible time-frequency structures is presented in a scene. Moreover, only parts of the data are relevant to discriminate between the different classes for sound scene or event classification tasks [17].

In neural network approach for acoustic signal classification, feature learning is usually conducted using basic features such as MFCCs or Mel-Filter Banks (MFBs). These features will be given as an input to a deep neural network such as CNN resulting in a feature map that can be used for classification.

Several studies [16], [18] have shown that higher classification accuracy can be achieved by aggregating features for environment sounds compared to single stream features. Multi-stream CNN networks are either based on decision level fusion or the fusion that occurs on the feature map layer before the classification layers. These models concatenate the outputs of the convolution layers from several CNNs and the concatenated feature maps are then used as inputs for either convolution layers or directly to the dense layers. In this study, we use a two-stream CNN architecture for feature aggregation to classify road surface wetness and is discussed in Section 3.

3 CNN ARCHITECTURES

Three CNN architectures are investigated for road surface wetness detection from recorded acoustic signals of tyre to road interaction. The first proposed CNN model maps a 2-dimensional input, in this case13 MFCCs, using several layers of convolution and fully connected layers to a probability vector over the two different classes of wet and dry. We refer to this architecture as MFCCs-CNN in this paper. The second architecture is a 1-dimensional version of

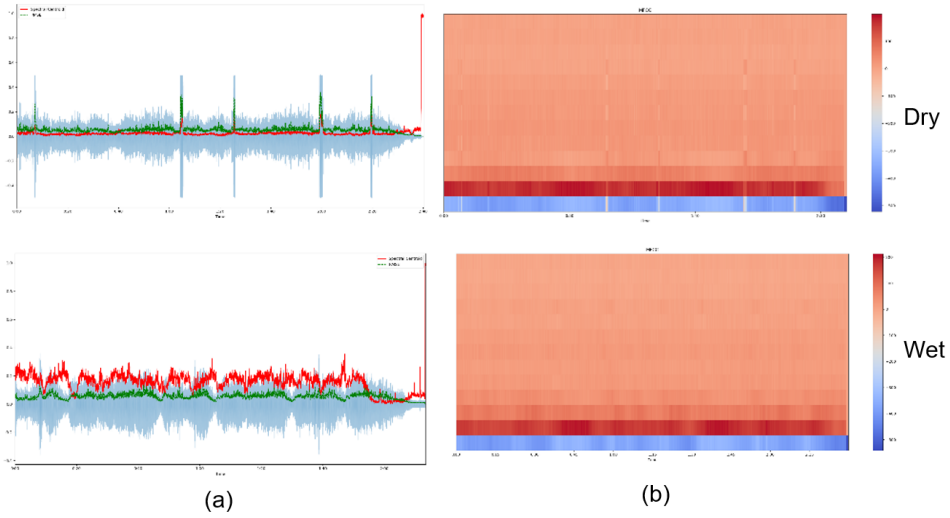

Figure 1: (a) waveform (blue), RMSE (green), spectral centroid (red) and (b) MFCCs visualization of recorded dry and wet road surface acoustic signals.

the MFCCs-CNN where it takes RZASSR features of the acoustic signal as the input. We refer to this architecture as Spectral-CNN in this paper. The third proposed architecture is a two-stream CNN which is resulted from merging the convolution layers of MFCCs-CNN and Spectral-CNN by concatenating the resulted feature maps of each stream. The architecture resulting from concatenation of MFCCs (M) and Spectral (S) CNNs is referred to as MS-CNN in this paper. The selection of kernel sizes and number of hidden units are based on the preliminary experiments conducted over 30 different combinations.

3.1 Single Stream CNN

3.1.1 MFCCs-CNN. The MFCCs-CNN is a 2-dimensional CNN architecture inspired by Alexnet [19] and ZFNet [20] where takes 13 MFCCs as the input. As discussed in [20] and [21], by placing pooling layers between convolution layers and using smaller kernel sizes, better network performances can be achieved. This also has been discussed to be more suitable for larger data sets. MFCCs-CNN architecture has two convolution layers, each interleaved with a pooling layer. The output of the convolution layers are followed by a drop out operation and deep fully connected layers for classification. the MFCCs-CNN architecture is shown in Figure 2. Kernel sizes in convolution layers and hyperparameters are selected as follows:

- The first layer of the network uses 32 filters, a kernel size of (5,5) and stride of (1,1). this is followed by (2,2) strided max-pooling and the activation function is ReLU.
- The second layer of the network uses 64 filters with a kernel size of (5,5) and stride of (1,1). this is followed by (2,2) strided max-pooling and the activation function is ReLU.
- The third layer is a fully-connected layer with 128 hidden units and the activation function is ReLU.
- The fourth layer is a fully-connected layer with 64 hidden units and the activation function is ReLU.
- The output is 2 units and the activation function is SoftMax.

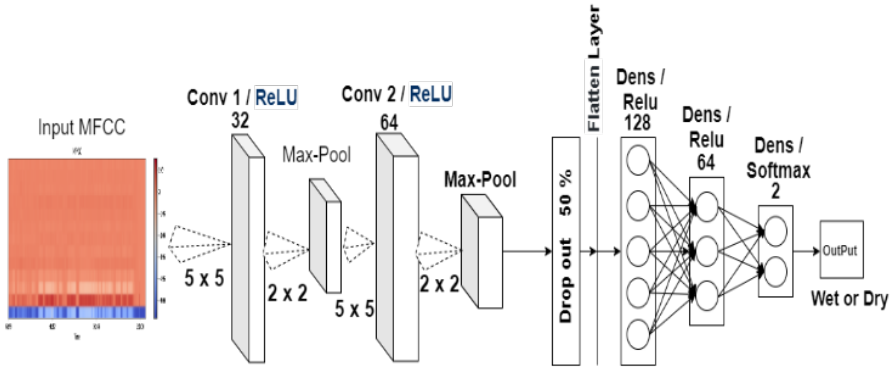

Figure 2: MFCCs-CNN network architecture.

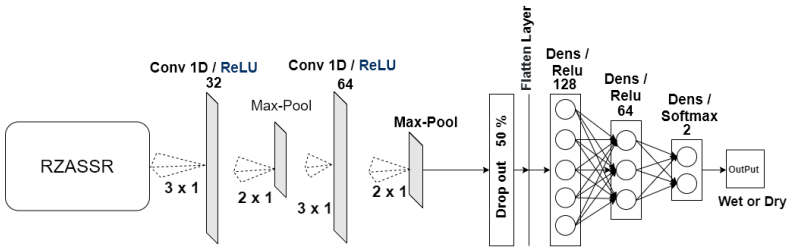

Figure 3: Spectral-CNN network architecture.

3.1.2 Spectral-CNN. The Spectral-CNN is a 1-dimensional version of the MFCCs-CNN with a similar network architecture as shown in Figure 3. The input for this architecture is a mix of time domain features and frequency domain features that are described in Section 2. The features that are used as input for the Spectral-CNN are RMS, ZCR, ASF, spectral contrast, spectral centroid and spectral roll-off. 1-dimensional CNN architecture is selected since the input features can only be presented in 1 dimension, unlike MFCCs which is a time-frequency domain and can be represented in 2-dimension.

3.2 Two-stream MS-CNN

By concatenating the resulted feature maps from the two single stream CNNs, MFCCs-CNN and Spectral-CNN, we propose a two-stream CNN architecture and we refer to it as MS-CNN for road surface classification.

The network consists of two separate CNN as shown in Figure 4. The CNN that learns spectral features, begins with 4 layers of convolution layers stacked together followed by a max-pooling layer. The second architecture that learns features from MFCCs begins with the convolution layer interleaved with a pooling layer, followed by fully connected layers at the end. Resulted feature map from the CNN models will be concatenated. To avoid overfitting a 50% drop out was used before the final dense layers.

The performance of the proposed two-stream CNN architecture is compared to the single stream MFCCs-CNN and Spectral-CNN architectures and discussed in Section 4.

4 EXPERIMENTATION AND RESULTS

For training all models, the optimize cross-entropy loss function and the Adam optimizer algorithm [22] was used. To adjust the hyperparameters of the Adam optimizer we followed the authors suggestion [22] and set the β_1 and β_2 which are first and second moment of the gradient to 0.9 and 0.999 respectively. For more details on Adam optimizer refer to [22]. To find the best learning rate, we tested the models over 10 epochs with learning rates of 0.1, 0.01 and 0.001 and compared the accuracy performance of the models. Learning rate of 0.001 was selected as it performed better compared to other learning rates across all models. Dropout of 50% is applied to the output of the max-pooling of each stream and then we flatten the data and pass it as an input to fully connected layers. The models are trained using an early stopping method, with a patience of 10 epochs and a maximum of 1000 epochs and a batch size of 128. These parameters are selected based on our hardware capability. Python and Keras had been used for the models implementation.

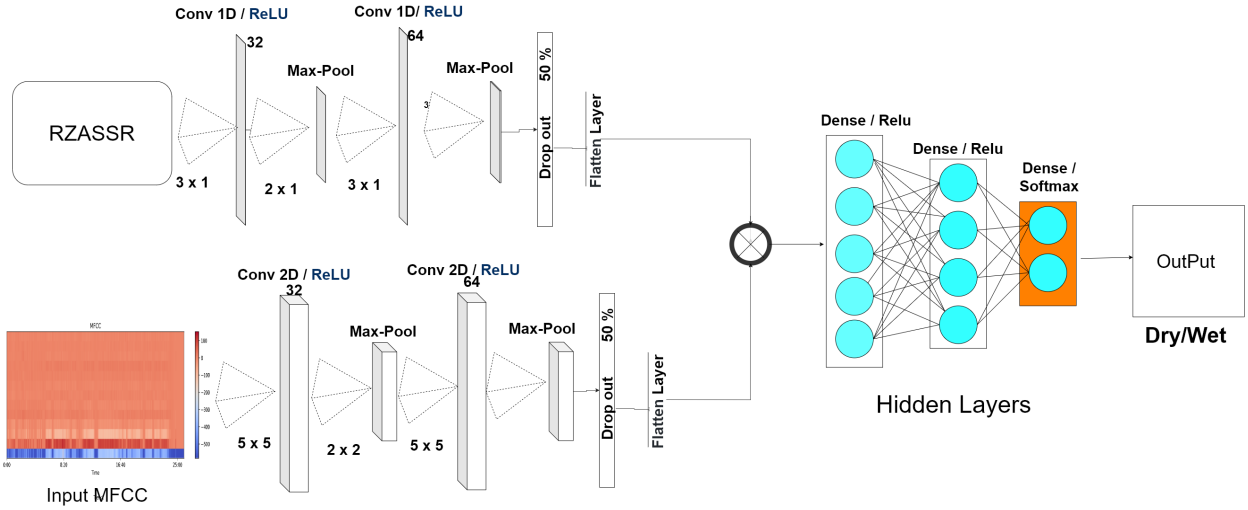

Figure 4: Two-stream MS-CNN architecture.

4.1 Dataset

The recording of tyre to road interaction acoustic signals have been conducted in a controlled environment. This way, environmental noises such as other vehicles on the road and construction noises are minimized and the water level on the road surface can be controlled.

The tyre to road interaction acoustic signals were recorded using Beyerdynamic MC 930 STEREO SET microphones, mounted at a close proximity of the tyre and the road surface as shown in Figure 5 (b). The previous works in the literature [5]– [8], [23], recorded acoustic signals from rear tyres farther from the engine and exhaust pipe using one microphone. In this study, the acoustic signals were recorded from two microphones placed on both left and right rear axis. Two different cars were experimented with. Car 1 is a 2012 Toyota Aygo and Car 2 is a 2015 VW Tiguan. In addition, a down facing camera on the side farther from the exhaust pipe and a dashcam mounted to the car's windshield was used to verify the road surface condition. The microphones were mounted to the front side of the tyres that are less affected by water splashing while driving on wet surfaces. A windscreen foam was installed on the microphones to lower the wind effect noise on the recorded audio. A Scarlett 2i4 (2nd Gen) USB audio recording interface connected to a laptop, controlled from the inside of the vehicle, was used to record the audio signals in uncompressed 24-bit depth format. For audio recording, Apple Garageband software installed on the in-cabin laptop was used. The input gain level was identified during preliminary experiments to obtain appropriate recording levels. Subsequently, the gain remained unchanged for all experiments. Figure 5 (a) shows the in-cabin interface setup.

(a) In-cabin setup (b) Microphone setup

Figure 5: Data acquisition setup (a) in-cabin laptop and interface(b) the left tyre microphone location.

A few preliminary experiments were initially conducted to calibrate and test the equipment such as finding the best gain level on the interface and best location to mount the microphones on each car. The tyre to road interaction data was collected in Millbrook Proving Ground, UK. 1mile straight and dynamics pad tracks which are both paved with asphalt have been selected for the recording sessions. Both tracks were recorded once in wet and once in dry condition. For wetting the surface of 1mile straight track a water bowser truck was used. Right before each recording session the truck drove on a constant speed and sprinkled water on the paved surface. To wet the surface of the dynamics pad track a water cannon was used. The water canon wetted a circular area of the track for 10 minutes before each recording session. A circular path with few turning points was driven through at a low speed to simulate urban driving on the dynamics pad. In total, we recorded 36 minutes of audio and video data on the Millbrook proving ground (the dataset is available for download at http://sbahrami.com/dataset/icimt20/).

4.2 Evaluation Metrics

To evaluate the performance of the models in addition to accuracy that was used by [14], recall, precision and F-measures are used [24]. The standard definitions adapted from [25] are as follows:

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{1}$$

$$recall = \frac{tp}{tp + fn} \tag{2}$$

$$precision = \frac{tp}{tp + fp} \tag{3}$$

$$F - measure = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{4}$$

Where tp is true positive, in this case wet road detected correctly, fp is false positive, in this case dry road detected correctly. The $\beta = 1$ is used to evenly balance the $-measure$, also refer to as F_1.

4.3 Results

The acoustic recordings were broken into 4 seconds clips to reduce the file size making it easier to process. To extract features discussed in Section 2, each clip is segmented into 30 milliseconds frames with a frame step of 10 milliseconds. Preliminary experiments with applying high-pass and linear phase filters to reduce the wind noises was conducted. By listening to the audio after applying these filters, the wind noise was reduced but it resulted in a 20% drop on the accuracy of the models, therefore filters were not applied. Each model is trained using early stopping with a patience of 10 epochs and for a maximum of 1000 epochs.

K-fold validations are generally used to overcome the overfitting problem. To make sure that we did not overfit the models 3-fold cross-validation was implemented. We divided all the samples into 3 parts and used 80% of the data for training and validation of the model and 20% for testing. The test data is only used for predication and kept the same during all 3-folds. We iterated 3 times over the 80% samples so that all the data were used for training and the validation of each model. The average mean accuracy of the models has been calculated based on each 3-fold.

Table 1 shows the classification performances for MFCCs-CNN, Spectral-CNN and MS-CNN architectures using differing feature sets when trained using data collected from both cars. The two-stream architecture has achieved the highest mean average accuracy of 92.29% clearly outperforming single stream architectures. It can be concluded that merging features learned from time and frequency domain with MFCCs improves the accuracy.

Further experiments were conducted to individually evaluate acoustic data collected from each of the cars, Car 1(C1) and Car 2(C2). We trained and tested the proposed CNN models using data

Table 1: Classification performance of two-stream compared to each stream individually.

Model	Feature Set	Accuracy	Precision	Recall	F_1
MFCCs-CNN	MFCCs	86.81%	86.95%	86.99%	86.81%
Spectral-CNN	RZASSR	69.63%	70.04%	67.06%	67.13%
MS-CNN	Concatenated	92.29%	92.28%	92.37%	92.28%

Table 2: Models performance with differing training and testing datasets.

	Spectral-CNN		MFCCs-CNN		MS-CNN	
Train/Test	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1
C1/C1	72.92%	72.87%	75.67%	75.56%	86.35%	86.32%
C1/C2	48.76%	47.88%	49.82%	48.39%	50.34%	49.50%
C2/C2	69.48%	69.29%	71.60%	71.60%	86.84%	86.83%
C2/C1	52.26%	52.24%	53.88%	53.66%	51.82%	51.76%

recorded from each of these cars individually to evaluate the effect of car and tyre type on the classification performance of the model. There has been studies [9] that investigated the impact of the tyre types, summer or winter tyre, on classification performance, but to best of our knowledge, there has not been any investigation on using different cars for data collection in the state of art. Table 2 Shows when the model is trained using C1 and tested using C2 as C1/C2 combination and inversely C2/C1 combination, the classification performance of the model drop by 25 % in MFCCs-CNN and up to 36% for MS-CNN model. This drop in classification performance can be related to tyre type and size of the two cars. Even though both cars were using summer tyres during the recordings, but Car 1 has smaller tyre compared to Car 2 (C2). The sound pattern difference of Car 1 and Car 2 recordings is visualized through mel-spectrogram in Figure 6. When the models are trained and tested using the acoustic recordings from one of the cars better accuracy performances can be achieved, for instance both C1/C1 and C2/C2 achieved 86% of accuracy on MS-CNN model. The performance of the models are slightly better when trained using Car 2 and tested against Car1. We believe this is due to the distance of the microphone to the engine on Car 2 which results in fewer engine noises in Car 2 recordings.

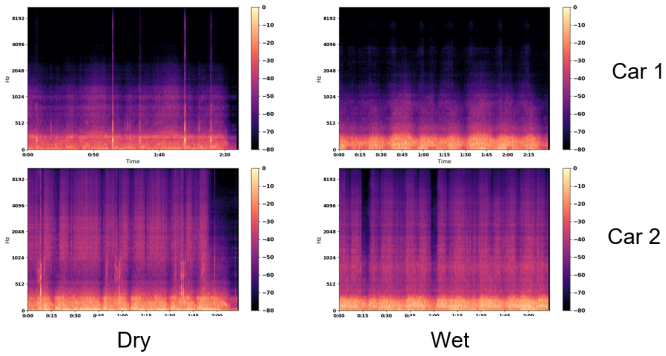

Figure 6: Mel-spectogram of dry and wet condition recorded on dynamics pad track.

5 CONCLUSION AND FUTURE WORK

In this study, we investigated a deep learning method for dry/wet road classification based on acoustic sensors mounted on a car. A two-stream CNN architecture was proposed and compared to single stream CNN architecture using differing feature sets for wet road surface detection. A dataset of tyre sound recording was compiled and used to train and evaluate the proposed models. MFCCs have been proven to be an effective feature set on its own for many audio classification tasks, however in this study we show the concatenation of MFCCs and RZASSR features outperform the accuracy performance of single stream architectures for road surface wetness detection. The accuracy performance of the two-stream MS-CNN architecture proposed is promising for road wetness classification. Further studies and experimentations with other time-frequency domain features such as constant-Q chromagram is required to develop a model that can detect road wetness across all car and tyre types. The current dataset would be expanded using more car models and urban routes. More robust evaluation using data augmentation techniques will be performed for road surface wetness detection and development of vehicle active safety systems.

ACKNOWLEDGMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 691154 STEP2DYNA and No 778602 ULTRACEPT.

REFERENCES

[1] "Annual Accident Report 2018," European road safety observatory, 2018. https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/asr2018.pdf (accessed May 29, 2020).
[2] Y. Q. Zhao, H. Q. Li, F. Lin, J. Wang, and X. W. Ji, "Estimation of Road Friction Coefficient in Different Road Conditions Based on Vehicle Braking Dynamics," Chinese J. Mech. Eng. (English Ed., vol. 30, no. 4, pp. 982–990, 2017, doi: 10.1007/s10033-017-0143-z.
[3] R. Ghandour, A. Victorino, M. Doumiati, and A. Charara, "Tire/road friction coefficient estimation applied to road safety," 18th Mediterr. Conf. Control Autom.

MED'10 - Conf. Proc., pp. 1485–1490, 2010, doi: 10.1109/MED.2010.5547840.
[4] W. Kongrattanaprasert, H. Nomura, T. Kamamura, and K. Ueda, "Automatic Detection of Road Surface States from Tire Noise Using Neural Network Analysis," Proc. 20th Int. Congr. Acoust., no. August, pp. 1–4, 2010.
[5] J. Alonso *et al.*, "On-board wet road surface identification using tyre/road noise and Support Vector Machines," Appl. Acoust., vol. 76, pp. 407–415, 2013, doi: 10.1016/j.apacoust.2013.09.011.
[6] P. Boyraz, "Acoustic road-type estimation for intelligent vehicle safety applications," Int. J. Veh. Saf., vol. 7, no. 2, p. 209, 2014, doi: 10.1504/IJVS.2014.060167.
[7] I. Abdic *et al.*, "Detecting road surface wetness from audio: A deep learning approach," in 2016 23rd International Conference on Pattern Recognition (ICPR), Dec. 2016, no. 1, pp. 3458–3463, doi: 10.1109/ICPR.2016.7900169.
[8] M. Kalliris, S. Kanarachos, R. Kotsakis, O. Haas, and M. Blundell, "Machine Learning Algorithms for Wet Road Surface Detection Using Acoustic Measurements," in Proceedings - 2019 IEEE International Conference on Mechatronics, ICM 2019, 2019, vol. 1, pp. 265–270, doi: 10.1109/ICMECH.2019.8722834.
[9] G. Pepe, L. Gabrielli, L. Ambrosini, S. Squartini, and L. Cattani, "Detecting road surface wetness using microphones and convolutional neural networks," 2019.
[10] P. Jonsson, J. Casselgren, and B. Thörnberg, "Road Surface Status Classification Using Spectral Analysis of NIR Camera Images," IEEE Sens. J., vol. 15, no. 3, pp. 1641–1656, 2015, doi: 10.1109/JSEN.2014.2364854.
[11] W. A. Cahyadi, Y. H. Kim, Y. H. Chung, and Z. Ghassemlooy, "Efficient road surface detection using visible light communication," Int. Conf. Ubiquitous Futur. Networks, ICUFN, vol. 2015-Augus, pp. 61–63, 2015, doi: 10.1109/ICUFN.2015.7182498.
[12] K. J. Piczak, "Environmental sound classification with convolutional neural networks," IEEE Int. Work. Mach. Learn. Signal Process. MLSP, vol. 2015-Novem, pp. 1–6, 2015, doi: 10.1109/MLSP.2015.7324337.
[13] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, "An Ensemble Stacked Convolutional Neural Network Model for Environmental Event Sound Recognition," Appl. Sci., vol. 8, no. 7, p. 1152, Jul. 2018, doi: 10.3390/app8071152.
[14] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Process. Lett., vol. 24, no. 3, pp. 279–283, 2017, doi: 10.1109/LSP.2017.2657381.
[15] Y. Shen, J. Cao, J. Wang, and Z. Yang, "Urban acoustic classification based on deep feature transfer learning," J. Franklin Inst., vol. 357, no. 1, pp. 667–686, 2020, doi: 10.1016/j.jfranklin.2019.10.014.
[16] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-Stream Network with Temporal Attention for Environmental Sound Classification," in Interspeech 2019, Sep. 2019, pp. 3604–3608, doi: 10.21437/Interspeech.2019-3019.
[17] R. Serizel, V. Bisot, S. Essid, and G. Richard, "Acoustic Features for Environmental Sound Analysis," in Computational Analysis of Sound Scenes and Events, Cham: Springer International Publishing, 2018, pp. 71–101.
[18] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion," Sensors, vol. 19, no. 7, p. 1733, Apr. 2019, doi: 10.3390/s19071733.
[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[20] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in 13th European Conference on Computer Vision, ECCV 2014, vol. 8689 LNCS, no. PART 1, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Zurich, Switzerland: Springer Verlag, 2014, pp. 818–833.
[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in International Conference on Learning Representations, 2014, pp. 1–14.
[22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Int. Conf. Learn. Represent., pp. 1–15, Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.6980.
[23] W. Kongrattanaprasert and T. Kamakura, "Detection of Road Surface States from Tire Noise Using Neural Network Analysis," IEEJ Trans. Ind. Appl., vol. 130, no. 7, pp. 920–925, 2010.
[24] L. Ambrosini, L. Gabrielli, F. Vesperini, S. Squartini, and L. Cattani, "Deep Neural Networks for Road Surface Roughness Classification from Acoustic Signals," Audio Eng. Soc. Conv., p. 9934, 2018.
[25] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation," in AI 2006: Advances in Artificial Intelligence, vol. 4304, no. 1, 2006, pp. 1015–1021.