Challenges of Vehicle Classification Using Acoustics

Raiden Evans, Stephanie Mullins, George Noel
Infoscitex, a DCS Company
revans@infoscitex.com, smullins@dcs corp.com, gnoel@infoscitex.com

Abstract
Automated acoustic classification of vehicles is a challenging problem with many variables. Vehicles produce complex sounds from many sources, sound signatures vary between similar vehicles, and background noise has a large impact on audio data. Past research studied different vehicle classification techniques but often relied on datasets with little variation in vehicle model, environmental conditions, or microphones. High-accuracy results on these datasets suggest issues of overfitting. This paper highlights the challenges of creating robust datasets for difficult acoustic classification tasks. Challenges include collecting on multiple instances of a target class, recording with varied environmental noise, and collecting across multiple types of microphones. The work presented also evaluates classification performance on combinations of acoustic feature sets and common machine learning algorithms. Classification F-score performance drops 38.67% when the test set has background noise that differs from the classifier’s training set. Performance also significantly drops when the classifier is tested on a vehicle of a model year not included in the training set. Lastly, the vehicle classifier heavily overfits on the signatures of the microphones it is trained on.

Introduction
Automated vehicle characterization has a wide range of potential uses, from security to law enforcement and traffic management. Vehicle acoustics provides another data modality, with each engine generating a unique signature. Previous research has successfully characterized traffic density and speed (Na et al. 2015) and some attempted to classify generic classes of vehicles (Mayvan et al. 2015) (Kubera, Wieczorkowska, and Skrzypiec 2015). Some research has demonstrated success in classifying specific vehicle models based solely on acoustics (Göksu 2018) (Guo et al., 2012); however, they were mainly limited to a small subset of vehicles in specific environments.

This paper explores acoustic vehicle classification using a range of vehicle types, feature sets, and classification algorithms. The primary goal of this research is to highlight the challenges of motor vehicle acoustics classification, from overfitting on environmental conditions to distinguishing between the same models across different years. The paper begins by describing the data collection procedures and experimental design, then compares feature sets and algorithms. It then demonstrates the challenges of classifying unseen data, especially when recorded on different days or using different instances of the same model/year.

Literature Review
Acoustic vehicle classification studies either focus on classifying between a few broad vehicle categories (e.g. car, truck, bus, motorbike) (Mayvan et al. 2015) or between a few specific vehicle types (Göksu 2018). George et al. (2013) create a neural network for distinguishing between heavy vehicles, medium vehicles, light vehicles, and horn noises. Sounds are collected on vehicles driving by the microphone. By using 30 mel-frequency cepstral coefficients (MFCC), an accuracy of 67% is obtained. The research conducted by Mayvan et al. (2015) classifies between buses, cars, motorbikes, and trucks. They use the feature extractors presented in Gorski and Zarzycki’s work (2012), which includes applying custom mel-filters to the audio. These custom features were found to classify at 80% accuracy between the four categories of vehicles. In Kubera, Wieczorkowska, and Skrzypiec (2015), 4-fold cross validation on an hour-long audio recording is performed to create a classification dataset for 5 vehicle categories. Using this data, they build a classifier to achieve accuracies of 99.93%, 68.79%, 81.48%, 74.07%, and 95.32% on small cars, vans, small trucks, big trucks, and tractors, respectively. Another paper demonstrates classification between cars, bikes, trucks, and lorries or large trucks (Paulraj et al., 2013). They are able to achieve an average accuracy of 94.5% by using a probabilistic neural network on coefficients of an auto-regressive model fit to each sound sample.

Some papers instead focus on classifying specific vehicle types. Only one vehicle instance is usually recorded to make...
up the data for a class, rather than collecting data on multiple instances of the same vehicle type. One paper classifies 5 vehicles at an average accuracy of 77.05% using key frequency features on a support vector machine (SVM) (Guo et al., 2012). Using a decision fusion approach, this accuracy is boosted to 86.06%. Jin et al. (2019) classify at 91.93% accuracy between two moving vehicles using seismic data. Features include mel-frequency cepstral coefficients (MFCC) and log-frequency cepstral coefficients (LFCC). The methods used with seismic data are very similar to those used with acoustic data. Aljaafreh and Dong (2010) find that wavelet packet decomposition slightly outperforms Short-time Fourier transform (STFT) coefficients and that SVM performed better than KNN in a binary classifier. Göksu (2018) claims that Fourier transformed time frequency features from wavelet packet decomposition provide stable results.

Past research reports a wide range of accuracies for acoustic vehicle classification. While each paper utilizes a unique classification algorithms and feature extraction methods.

### Data Collection

To measure how well specific vehicle makes and models can be distinguished using acoustic data, a variety of consumer vehicles are recorded in a range of operating conditions. Engine audio is recorded during idle and stationary revving with both the hood up and the hood down. To collect engine revving sounds, the vehicle is shifted into park, and the engine is revved to around 2,000 RPM. During a recording session, about 40 seconds of 2,000 RPM revving is recorded and 10 seconds of 3,000 RPM revving is recorded. Microphones are placed in front of the vehicle for idle and revving recordings. The first and last 5 seconds of each sound sample is deleted to ensure clean data.

Additionally, the vehicles are recorded driving by the microphones at a roughly constant speed of 15 miles per hour. Microphones are placed next to one another in a line parallel to the driving direction about 5 feet from the passing side of the vehicle. Each drive-by audio clip is manually trimmed to contain only audio of the vehicle and the average duration of a drive-by audio clip is about 3 seconds. The vehicles that comprise the dataset are shown in Table 1.

Acoustic data is collected on vehicles using five different microphones, including a Rode VideoMic and an Audio-Technica AT2020. Also included are the microphones on a Samsung Galaxy s8, a Samsung Galaxy s20, and an Apple iPhone 11 Pro. Audio is recorded at 44.1 kHz. The iPhone records at 48 kHz, so this audio is resampled to 44.1 kHz to maintain consistency across the data.

| Vehicle Name                      | Vehicle ID | Day # | Total Seconds Recorded |
|-----------------------------------|------------|-------|------------------------|
| Chevy Silverado 2007              | CS-07      | x     | x                      | 1,181 |
| Chevy Silverado 2013              | CS-13      | x     | x                      | 1,065 |
| Dodge Ram 1500 2015               | DR-15      | x     | x                      | 805   |
| Ford F350 2020                    | F350-20    | x     | x                      | 1,645 |
| Ford Transit 2016 (1)             | FT-16-1    | x     | x                      | 816   |
| Ford Transit 2016 (2)             | FT-16-2    | x     | x                      | 883   |
| Ford Transit 2020                 | FT-20      | x     | x                      | 814   |
| Honda Accord 2017                 | HA-17      | x     | x                      | 840   |
| Honda Civic 2010                  | HC-10      | x     | x                      | 1,130 |
| Honda Civic 2016                  | HC-16      | x     | x                      | 2,100 |
| Hyundai Elantra 2020 (1)          | HE-20-1    | x     | x                      | 1,706 |
| Hyundai Elantra 2020 (2)          | HE-20-2    | x     | x                      | 1,647 |
| Hyundai Elantra 2021              | HE-21      | x     | x                      | 831   |
| Hyundai Sonata 2019               | HS-19      | x     | x                      | 818   |
| Mercedes CLK 320 2001             | M320-01    | x     | x                      | 895   |
| Subaru Legacy 2016                | SL-16      | x     | x                      | 2,137 |

Table 1. Vehicle Collection Day Matrix
While recording audio, environmental acoustic noise is minimized as much as possible by timing the recordings during quiet moments. However, background noise is not completely eliminated. Some environmental factors include wind, lawnmowers, other vehicles in the area, and idling aircraft (some recording takes place near a tarmac).

On days 3 through 5, some microphones were placed to the side of the vehicle for idle and revving recordings. This was performed to test how positional data variation affected algorithm performance, from which no significant results were found. Microphones placed to the side of the vehicles (for idle and revving recordings) are as follows: Day 3, Rode and Galaxy s8 mics; Day 4, Galaxy s20 mic; Day 5, Galaxy s8 mic.

**Experimental Procedures**

In all the experiments, training and testing data is generated by dividing the recorded audio into distinct quarter-second intervals with no overlap. Each chunk of audio is preprocessed and features are extracted. Unless otherwise stated, 5-fold cross-validation is performed using randomized shuffle to calculate performance. By default, each vehicle make and model is treated as a distinct class, yielding a total of 10 classes. Some classes contain vehicles manufactured in different years. Because drive-by data could not be recorded on day 5, the Subaru Legacy and Dodge Ram 1500 do not have any associated drive-by data. Drive-by classification, therefore, includes eight classes instead of ten.

F-score is chosen as the reporting metric to compensate for class imbalance. For a given experiment, the final reporting score is the equally-weighted average across all classes’ F-scores.

An initial experiment is conducted to identify the combination of classification algorithm and feature extraction method with the highest F-score. The best performing classifier-feature combination is then used to measure performance on the remaining experiments.

K-nearest neighbors (KNN), support vector machine (SVM), and random forest classifier (RFC) algorithms are compared. Algorithm hyperparameters are tuned, including the kernel function used in the SVM.

Feature sets include the Short-time Fourier transform (STFT), mel-frequency cepstral coefficients (MFCCs), and gammatone frequency cepstral coefficients (GFCCs). The dimensionality of each feature is tuned. Dimensionality is reduced on the STFT using an evenly-spaced filter bank. The last feature set uses the power of the final nodes of wavelet packet decomposition (WPD) to build feature vectors. WPD dimensionality is tuned by changing the number of decomposition levels. The coiflet 6 wavelet is found to work well for the vehicle data collected.

The primary libraries used for classification algorithms and audio feature extraction include scikit-learn, librosa, pywt, and pyAudioProcessing.

**Experiments**

Five experiments are conducted to help characterize the data. These generally start with simple classification problems with complexity added in later experiments.

Experiment 1: Each algorithm and feature set combination is trained and tested on drive-by data to determine the best combination. This combination is then used for subsequent experiments. The drive-by data was chosen because initial experiments demonstrated lower classifier accuracy with this data set. This helped distinguish between classifier performances.

Experiment 2: Performances are evaluated on all recorded audio from each vehicle model, including idling, revving and drive-by acoustics. Baseline performance is found for each of these three operating conditions to compare classification difficulty.

Experiment 3: This experiment trains on two vehicles of the same model and tests on a third. The Hyundai Elantra and Ford Transit both included three different vehicles of the same model. In each case, at least one vehicle was a different year than the others. This experiment tests whether there is a difference between the same make and model with different years.

Experiment 4: Environmental sounds and other possible variables varied from day to day. This experiment trains a classifier on three days of data and tests on an unseen fourth day. This experiment provides evidence on how much the classification model is training on day-specific information.

Experiment 5: Data is collected using both high-quality microphones and smartphone microphones. This experiment measures the performance of a model when trained...
on only high-quality microphone data versus a model trained on only smartphone microphone data. Additionally, performance is measured when a model is trained on high-quality microphone data and tested on smartphone microphone data, and vice versa.

**Results**

For the following experimental results, all comparisons are tested for significance using a paired bootstrap test with p=0.05. Any lack of significance is indicated in the text.

Experiment 1 compares SVM, RFC, and KNN algorithms. Parameters are tuned to find the best performance, with k=2 found for KNN. RFC works best when each decision node considers 30% of the feature dimensions. 1,000 decision trees are used, with each decision tree fitting to the full set of training data. A 3rd degree polynomial kernel works best for the SVM. The experiment finds that a large regularization parameter of 1,000 produces the highest F-score for the SVM. A parameter this large usually leads to overfitting, so the SVM is tuned on additional experiments. No difference between larger and smaller regularization parameters is found on these experiments. Despite its large magnitude, the regularization parameter does not increase overfitting on the data. Both SVMs and RFCs performed well on drive-by data while KNNs did not classify as well. Both the RFC and SVM models are evaluated in later experiments, but only the RFC performance is reported as it is found to be slightly more consistent.

| Table 2. Experiment 1: Classifier-Feature Combination F-Score Performance |
|---------------|----------------|----------------|----------------|
|               | FFT    | MFCC | GFCC  | WPD   |
| KNN           | 63.28% | 62.89% | 72.50% | 42.99% |
| SVM           | 66.93% | **81.02%** | 74.52% | 52.46% |
| Random Forest | 70.63% | **80.24%** | 67.19% | 59.62% |

Parameter tuning is also performed on all feature extractors. Both MFCCs and GFCCs perform best when 30 coefficients are used. STFT coefficients classify best when each quarter-second chunk of sound is calculated using a centered 16,384 sample window with reflection padding. These coefficients are then reduced to 200 coefficients using a series of evenly-spaced triangular filters. WPD performs best using 8 levels of decomposition, resulting in 256 features.

Experiment 2 calculates a baseline F-score of the different data collection recordings (idle, rev, idle and rev, and drive-by). This experiment includes days three, four, five, and six from Table 1 because the recording process and the placement of the microphones are the most consistent with each other. Note that no drive-by data is recorded for day five. The results, located in Table 3, show that most have near perfect scores with performance degrading in the drive-by test.

Experiment 3 tests vehicles of the same make and model but of different years. Two tests are conducted within this experiment. For each, all ten models of vehicles are used to train the classifier in the idle and rev task. All eight models of vehicles are used for the drive-by task. However, a single Ford Fusion and a single Hyundai Elantra are left out of the

| Table 3. Experiment 2: Idle, Rev, and Drive-by F-Score Performance |
|-----------------|
| F-score         |
| Idle            | 99.96% |
| Rev             | 99.62% |
| Idle + Rev      | 99.81% |
| Drive-by        | 79.75% |

| Table 4. Experiment 3: Dataset Matrix |
|------------------------------------|
| Train with Different Years, Test with One of the Same Year | Train with Same Years, Test with Different Year |
| Train | Test | Train | Test |
| HE-20-1 | X | X |
| HE-20-2 | X | X |
| HE-21 | X | X |
| FT-16-1 | X | X |
| FT-16-2 | X | X |
| FT-20 | X | X |
| All other vehicles | X | X |

| Table 5. Experiment 3: Train on 2 Vehicles, Test on 3rd Vehicle F-Score Performance |
|------------------------------------|
| Train with Different Years, Test with One of the Same Year | Train with Same Years, Test with Different Year |
| Idle + Rev Classification | 87.28% | 52.98% |
| Drive-by Classification | 88.93% | 75.81% |
training data and used to test the classifier. In the first test, classifiers are trained on the same year vehicle as in the test set. In the second test, the classifier training data did not include the test vehicle’s manufacturing year. See Table 4 for a breakdown of the vehicles used. The F-scores reported in Table 5 indicate that the classifier has difficulty when it has not been trained on the specific year being tested.

Experiment 4 measures how day-specific factors impact model performance. Acoustics collected on three separate days are used as the training set while an unseen fourth day serves as the test set. Note that only two vehicles are recorded over all four days, so the test set consists only of audio from the Honda Civic and Mercedes CLK320. See Table 6 for clarification on the training set and test set split. Performance significantly drops when testing on an unseen day’s data as seen by comparing Table 7 with Table 3.

The objective of experiment 5 is to determine if having different quality microphones can change the F-scores significantly. No statistically significant difference from the baseline F-scores is found when a model is trained and tested solely on high-quality microphone data, with F-scores of 99.83% for idle and revving and 80.27% for drive-by. This is similar to when a model is trained and tested on only smartphone microphone data, with F-scores of 99.76% for idle and revving and 80.31% for drive-by. A significant drop in performance is found, however, when the microphone type for the training data is different from the testing data. When the model is trained on the AT2020 and Rode and tested on smartphones, the F-scores are 22.46% for idle and revving and 18.57% for drive-by.

When the model is trained on the smartphones and tested on the AT2020 and Rode, the F-scores are 27.17% for idle and revving and 22.50% for drive-by. These results can also be found in Table 8.

### Discussion

As with similar research (George et al. 2013), this research found that MFCCs work the best for classifying a wide variety of vehicle makes and models when compared against the other features tested in this paper. Most existing vehicle acoustics classification research used either neural networks or SVMs. This research also explored SVMs and found that RFCs yield comparable performance.

The key finding from this research is how impactful microphones, environmental factors, and small vehicle differences can be in automated classification. A classifier will train on wind or background noise if data collection is not varied enough to account for these variables. Experiment 2 classified with exceptionally high F-scores since the algorithms were trained on data with the same background noises and equipment as the test data set. Performance dropped off significantly when data from another day or an “unseen” vehicle was used as the test data. That is not to say it performed poorly—the results were still far better than chance. However, it highlights the risk of overselling algorithm performance if experiments in acoustics classification are not designed properly.

Classifying vehicle by model also has to deal with variations between manufacturing years. Manufacturers often vary engine characteristics significantly, changing the acoustics enough to make classification difficult. Variations with microphones exacerbate the issue, highlighting the challenges with characterizing vehicles based on acoustics.

This research attempted to control many variables during data collection. These included microphone distance to the target and ensuring minimal ambient noise. A real-world environment will likely have significant background noise and need to classify vehicles from a variety of distances. As the algorithms in this research likely trained on feature amplitude, it may not perform well with variations in distance or
lower signal to noise ratios. Normalizing the feature vectors might compensate for this issue.

Conclusions and Future Work

This research finds that high F-scores are achieved in classifying vehicle makes and models when the test set is sampled randomly from the original data. Experiments indicate that collecting on a wide variety of make and model years is critical for building a robust classifier. Additionally, data needs to contain a wide variety of background noise to prevent overfitting and skewing of theoretical test results. While a model built and tested on smartphone microphone data has comparable results with one built and tested on high-quality microphones, it is crucial to build the dataset using microphones intended for final use. This may require using a variety of microphones or artificially varying the training data to better generalize the models.

This research builds an important foundation for future work in Cyber Physical Fused Sensing (CPFS). CPFS exploits the effect induced by the physical world into the cyber-domain. The effect itself, which only exists as a product of cyber-systems, can be amplified by fusing data from multiple modalities, such as microphones and video. Future work aims to further develop vehicle acoustic algorithms for a CPFS framework.

Next steps in this research will explore solutions to overfitting using data augmentation approaches. One common approach is to overlay background noise on top of the original data to lower the signal to noise ratio and provide a wide distribution of environmental sounds. Another interesting data augmentation approach is frequency warping, where the data is resampled at a different time scale. This technique shows promise in a paper on acoustic drone detection and classification (Kolamunna et al. 2021) and may provide robustness for vehicles moving at high speed.

Acoustic analysis provides an additional modality when characterizing vehicle behavior. This may involve detecting modifications to a vehicle’s base configuration or detecting overloaded vehicles. Future work in early engine failure detection could build upon the work presented by Wang et al. (2020) to identity issues before a vehicle is disabled. These are, however, complex inferences and future research must address the issues discussed in this paper if it is to be successful.

Acknowledgements

This work was funded under AFRL Contract Number: FA8650-18-C-1669 and is approved for public release; distribution unlimited: AFRL-2022-0194

References

Aljaafreh, A. and Dong, L. 2010. An Evaluation of Feature Extraction Methods for Vehicle Classification Based on Acoustic Signals. 2010 International Conference on Networking, Sensing and Control (ICNSC), 570-575. doi.org/10.1109/ICNSC.2010.5461596

George, J.; Cyril, A.; I. Koshy, B. I.; and Mary, L. 2013. Exploring Sound Signature for Vehicle Detection and Classification Using ANN. International Journal on Soft Computing, 4(2), 29-36. doi.org/10.5121/ijsc.2013.4203

Göksu, H. 2018. Engine Speed-Independent Acoustic Signature for Vehicles. Measurement and Control (United Kingdom), 51(3-4), 94-103. doi.10.1177/0020294018769080

Gorski, M. and Zarzycki, J. 2012. Feature extraction in vehicle classification. 2012 International Conference on Signals and Electronic Systems, 1-6. doi.10.1109/ICSES.2012.6382264

Guo, B.; Nixon, M. S.; and Damarla, T. 2012. Improving acoustic vehicle classification by information fusion. Pattern Analysis and Applications, 15(1), 29-43. doi.10.1007/s10044-011-0202-5

Jin, G.; Ye, B.; Wu, Y.; and Qu, F. 2019. Vehicle Classification Based on Seismic Signatures Using Convolutional Neural Network. IEEE Geoscience and Remote Sensing Letters, 16(4), 628–632. doi.org/10.1109/LGRS.2018.2879687

Kolamunna, H.; Dahanayaka, T.; Li, J.; Seneviratne, S.; Thilakaratne, K.; Zomaya, A. Y.; and Seneviratne, A. 2021. DronePrint: Acoustic Signatures for Open-set Drone Detection and Identification with Online Data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(1), 20:1-20:31. doi.org/10.1145/3448115

Kubera, E.; Wieczordowska, A.; and Skrzypiec, K. 2015. Audio-Based Hierarchic Vehicle Classification for Intelligent Transportation Systems, Vol. 938, 343-352. doi.org/10.1007/978-3-319-25252-0_37

Mayvan, A. D.; Beheshiti, S. A.; and Massom, M. H. 2015. Classification of Vehicles Based on Audio Signals using Quadratic Discriminant Analysis and High Energy Feature Vectors. International Journal on Soft Computing, 6(1), 53-64. doi.org/10.5121/ijsc.2015.6105

Na, Y.; Guo, Y.; Fu, Q.; and Yan, Y. 2015. An acoustic traffic monitoring system: Design and implementation. 2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing, 119–126. doi.org/10.1109/UICT-ScalCom-CBDCom-IoP.2015.41

Paulraj, M. P.; Adom, A. H.; Sundararaj, S.; and Rahim, N. B. A. 2013. Moving vehicle recognition and classification based on time domain approach. Procedia Engineering, 53, 405-410. doi.org/10.1016/j.proeng.2013.02.053

Wang, Y. S.; Liu, N. N.; Guo, H.; and Wang, X. L. 2020. An engine-fault-diagnosis system based on sound intensity analysis and wavelet packet pre-processing neural network. Engineering Applications of Artificial Intelligence, 94(333), 103765 1-11. doi.org/10.1016/j.engappai.2020.103765