Audio Classification of Accelerating Vehicles

An-Chih Yang* Emmett D. Goodman*

December 14, 2019

Abstract

Many emerging technologies depend on effective vehicle recognition to understand the evolution of traffic patterns, predict individual vehicle behaviors, and monitor roadway usage. In vehicle identification and classification, computer vision and lidar-based technologies dominate, although these approaches involve extensive computing power or expensive detection schemes. Audio identification of vehicles is promising because it requires simple and cheap recording devices and much lower quantities of data than other technologies. Importantly, audio-based technologies don't suffer from low visibility and are equally effective in low-light conditions. In this work, using careful feature extraction and various deep learning architectures, we demonstrate that vehicles can be effectively classified by recording vehicle acceleration with a cellphone.

1 Introduction

Vehicle identification and classification is an important challenge with applications ranging from autonomous driving to traffic management. Typically, vehicle characterization is done using image processing techniques which require significant computational power carrying out calculations in real-time [1]. Other light-based classification approaches including infrared (IR) [2] and radiowave (Radar) technologies [3], or even magnetic approaches [4], have be studied to supplement visual vehicle classification. Acoustic, or sound-based classification schemes provide important benefits compared to the more common light-based classification schemes. First, audio-based classification schemes would perform well in low light and poor weather conditions, where light-based technologies (i.e. ComputerVision/Radar/IR) would fail. Second, audio-based systems are small, cheap, and do not interfere with the flow of traffic. Finally, audio-based characterization schemes may provide critical information on important vehicle properties, such as engine or tire conditions, which could be key in military applications or in diagnosing vehicle condition.

Although some work has been done in audio-based vehicle classification, studies target classification of vehicles moving at constant velocity. Furthermore, these works are generally limited to classification between 3-4 main vehicle classes. An important and novel contribution of our work is the classification of vehicles accelerating from a full stop. Here, classification of accelerating vehicles offers unique challenges and opportunities. Importantly, vehicle acceleration will sound different depending to specific driving habits. However, compared to constant-velocity audio-data, acceleration provides rich information about the engine dynamics of the vehicle. For example, at highway speeds, a hybrid and a pickup truck may sound similar, but while in acceleration from a stop, the high wine of an electric motor will be distinct from the deeper vibrations of a diesel engine. In future work, audio features may allow clearer distinction between specific auto manufacturers compared to light-based classification technologies.

2 Related Work

Significant work has been done in classifying vehicles operating at constant velocity highway speeds. Across works, authors comment on challenges of data acquisition, including simultaneously arrival of numerous

*Stanford University, Palo Alto. {egoodma,angely}@stanford.edu.

vehicles, and the confounding effects of vehicle acceleration. Additionally, past works emphasize the importance of careful feature selection. In the late 90s, Nooralahayan first attempted to classify vehicles via audio recorders, and achieved an impressive 82 % accuracy between motorcycles, light vehicles, and buses [5]. Later, George et al. studied classification between three types of vehicles (light, medium, and heavy), and achieved a 67 % classification accuracy using artificial neural networks (ANN) [6]. In this case, the authors using smoothed audio intensities to generate key features. Dalir et al. found that by using various features of the audio data, including energy intensity and zero-cross rate, they could classify buses, cars, motors, and trucks, with up to 80 % accuracy using a support vector machine with Mel filter coefficients. In another three-class problem, Alexandre et al. carefully studied the feature-selection process for vehicle classification using a genetic algorithm with restricted search paired with an Extreme Learning Machine (ELM). Interestingly, by using this approach to hone in on a subset of features, the researchers improved classification accuracy from 75% to 94% [7]. Finally, in a recent paper Wieczorkowska et al. were able to achieve 74% accuracy in classification between 7 classes of vehicles, including buses, small trucks, big trucks, vans, motorcycles, cars, and tractors using a deep-learning model with two hidden layers of fifty neurons each with regularization and dropout [8]. These examples provide important insight into how to characterize vehicle audio properties, which we aim to translate into the accelerating vehicle problem.

3 Dataset and Features

3.1 Data Collection

Vehicle acceleration was collected and labeled by the authors. To avoid the confounding effects of oncoming traffic and simultaneously arriving vehicles noted in previous works, audio was recorded at an isolated stop sign. A cellphone (Samsung S9+) was fastened to a light-post directly adjacent to the stop sign, allowing direct vehicle visibility and clear audio recording of the accelerating vehicle. In total, 936 audio clips were extracted from 5 hrs of video, and divided into 7 classes and numerous manufacturers. Commercial vehicles represents the smallest classes, where only 40 training examples were acquired **(Figures 1,2)**.

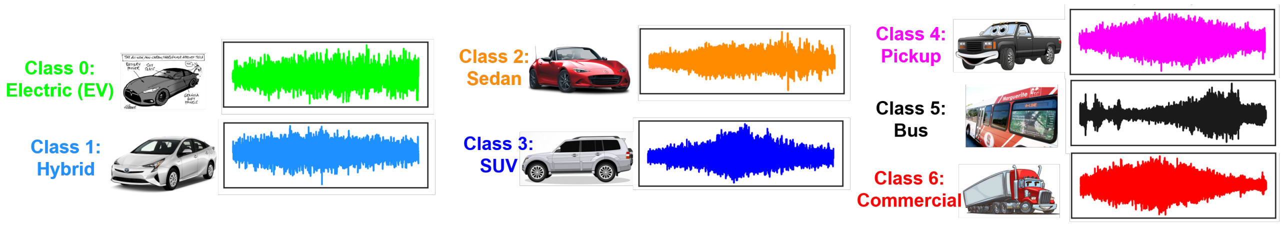

Figure 1: Representative non-normalized waveforms of seven vehicle classes studied in this work.

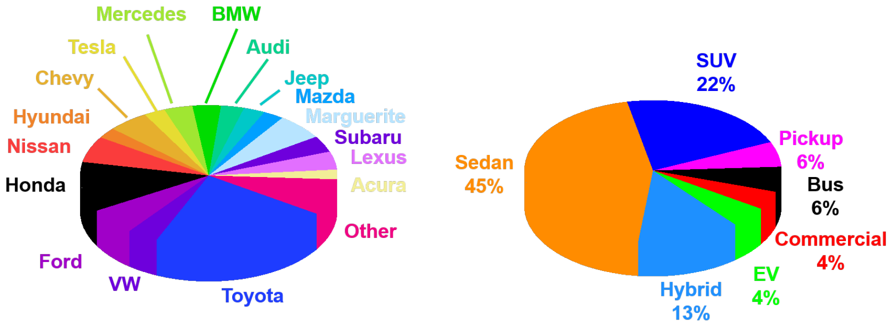

Figure 2: Manufacturer and class breakdown of collected data.

3.2 Feature Extraction

Temporal (i.e. time domain) and spectral (i.e. frequency domain) features were extracted from the audio acquired via Samsung Galaxy S9+ .mp4 files. The audio track was directly imported into Audacity software for splitting and labeling, and no further audio processing was performed prior to analysis. Labeling was performed by matching the audio track to the simultaneously taken video. In the temporal domain, 53 features were extracted including intensity variance, dynamic range, and max intensity (DR, I_{max}, I_{var}). Additionally, the audio track was evenly divided into 50 sections which were processed to reflect max intensity within each section. The spectra was then transformed into the frequency domain via the Fast-Fourier Transform (FFT) to access spectral features. The FFT was evenly divided into twenty-five sections, which were processed to extract the maximum intensity from each section, as in the temporal domain. The ratio of each spectra feature to each other was taken to produce a matrix of $25 \times 25 = 625$ spectral features. Combining 53 temporal and 625 spectral features, our classification scheme involved a total of 678 total features **(Figure 3)**.

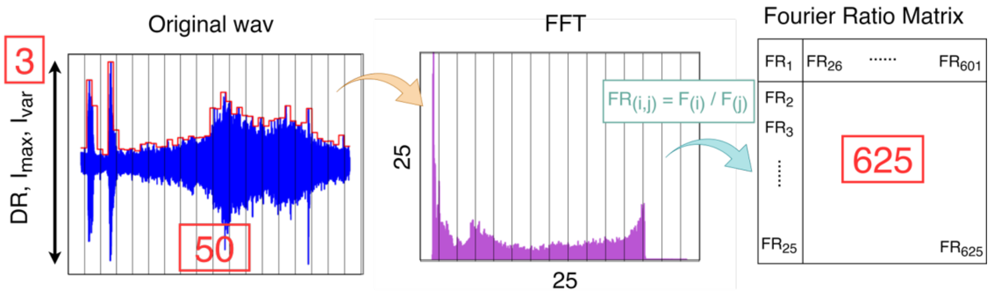

Figure 3: Feature extraction in both temporal and spectral spaces.

3.3 Data Augmentation

Data augmentation was used to create additional training examples for sparse classes. Here, both noise injection, and change of pitch, were used to create 1160 additional examples for classes that had less data, which were only added to the training set. With the augmented data, the 7 classes are more balanced and number of total examples increased to 2087 **(Figure 4)**.

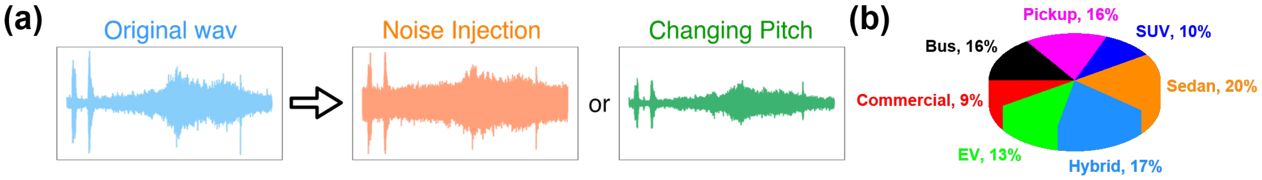

Figure 4: (a) Brief overview of data augmentation before/after noise injection and changing pitch. Example data taken from class 5 (Bus). (b) Class distribution after data augmentation.

4 Models

4.1 Fully-connected Neural Networks (FCNN)

For the FCNN models, either no hidden layers or one hidden layer with 20 neurons (with $ReLU$ as the activation function) were studied. Training examples with 678 features generated from the raw .wav file and FFT were sent to the model and batch gradient descent was used in the backpropagation process. Cross-Entropy Loss ($J_{CE} = \sum(y \log \hat{y})$) and Adam were used as the loss function and optimizer, respectively. In the output layer, we used $softmax$ as the activation function to determine the result for the full 5-class and 7-class problems **(Figure 5a)**. To minimize problems related to class imbalance, training and test examples were taken such that the most populated class (sedans) had no more than two times as many examples as the least populated class (commercial).

4.2 Recurrent Neural Networks (RNN)

For the RNN models, each neuron has *tanh* and *softmax* as the activation functions. Batch gradient descent was used in the backpropagation process. Again, Cross-Entropy Loss ($J_{CE} = \sum(y \log \hat{y})$) and Adam were used as the loss function and optimizer, respectively. Since RNN is a recursive method, we only want to feed in features which are in the form of time-series. Therefore, we feed in our raw .wav file which is divided into 50 groups. In each group, 50 features were determined by again dividing the group into 50 sub-groups and the max .wav signal in each sub-group was used to be the representative feature for the sub-group. Eventually, 50 × 50 features were generated and each group was sent into the RNN neuron once at a time. In the last (the 50^{th}) neuron, the class of the examples was predicted compared with the real label **(Figure 5b)**.

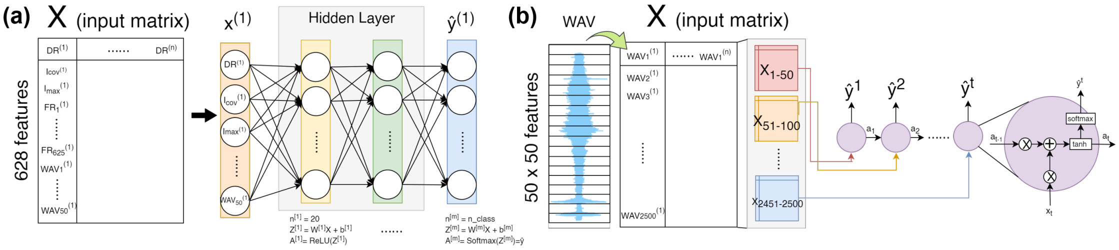

Figure 5: Scheme of our (a) FCNN model and (b) RNN model.

5 Results and Discussion

For both FCNN and RNN models hyperparameters were varied to increase test accuracy, including the learning rate and the number of neurons and hidden layers. For learning rate, 0.1, 0.03, 0.01, 0.003 and 0.001 were used, and our model iterated 30,000 times. The loss value converged for both training set and test set at each learning rate. For most models, we observed the general trend that as learning rate decreases, higher accuracy is achieved, and the curve of loss-iteration becomes flatter, probably because higher learning rate results in an unstable approach to the global/local minimum. For the tuning of the number of hidden layers and neurons, it is observed that there is no significant difference on accuracy between the number of hidden layers and neurons. For that reason, we used either no hidden layer or just 1 hidden layer with 20 neurons in our FCNN model.

Different sets of features were used as input for a variety of different deep learning architectures. In many cases, we used the entire suite of 678 temporal and spectral features. Additionally, using pairwise support vector machine analysis (not discussed here), we curated a smaller subset of 40 features that most effectively separated classes pairwise. With 40 features, we restricted our feature space in an aim to avoid overfitting. For these sets of features and models, we tried to tackle the 5-class and 7-class problems, and the results are shown in the table below:

No.	# Input Features	Model	Accuracy for 5-class	Accuracy for 7-class
1	678	FCNN, no hidden layers	Train: 94%; Test: 69%	Train: 79%; Test: 38%
2	678	FCNN, 1 hidden layer	Train: 91%; Test: 65%	Train: 69%; Test: 43%
3	40	FCNN, no hidden layers	Train: 94%; Test: 75%	Train: 82%; Test: 36%
4	40	FCNN, 1 hidden layer	Train: 97%; Test: 66%	Train: 64%; Test: 52%
5	678	FCNN, 1 hidden layer, w/ augmented data	Train: 91%; Test: 64%	Train: 91%; Test: 61%
6	50 ×50	RNN, w/ augmented data	Train: 82%; Test: 65%	Train: 62%; Test: 54%

Overall, we observed significantly better classification accuracy when distinguishing between five vehicle classes rather than seven vehicle classes. When distinguishing between five classes, we observed that models performed better than using no hidden layers, and when trained on a subset of the features (i.e. 40 rather than 678). By selecting specific features, we are effectively regularizing our model which allows us to achieve higher generality and accuracy on our test data. Training and test accuracies were approximately 20% lower when distinguishing between seven classes of vehicles. In this more complex system, we found the addition of a hidden layer increased the test accuracy, while the more selective set of parameters maintained a higher test accuracy still. It is possible that when distinguishing between more classes, having an additional hidden layer allows effective classification between very similar classes, such as electric vehicles and hybrids.

In many cases, it was difficult to distinguish between classes with very similar features. To visualize this challenge, we took the best models for both 5-class and 7-class problems and plotted them in the form of confusion matrices **(Figure 6)**. For the 7-class problem, we can see our model is doing a good job identifying class 4 (Pickup), class 5 (Bus) and class 6 (Commercial) with >80 % accuracy. Class 0 (EV) and class 1 (Hybrid) have 76% accuracy and most of the wrong predictions fall into the other class because of their similar features. The same issue is observed on the classification between class 2 (Sedan) and class 3 (SUV), likely because most sedans and SUVs are composed of similar engines. It is noted that no car in class 3 is successfully identified because of the varied features which could be placed into different classes.

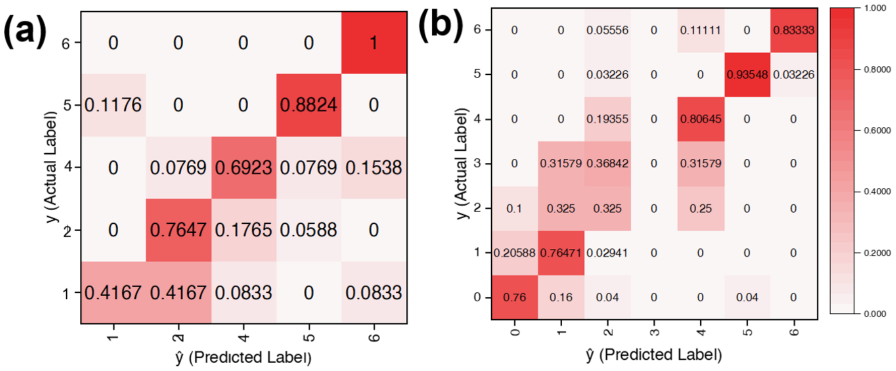

Figure 6: Confusion matrices of best model in (a) 5-class and (b) 7-class classification problems.

We also used an RNN model for both 5-class and 7-class classification problems. With respective accuracies of 65% and 54% on the 5-class and 7-class problems, we find that the RNN model does not outperform the FCNN model. We believe it is because our current RNN model does not involve any long-term memory of data in the early stage, while in certain vehicle classes, early-stage features are likely important. To improve this, Long-short Term Memory (LSTM) models will be used in the future.

6 Conclusions and Future Work

Overall, we found that 5 to 7 vehicle classes can be successfully distinguished using cellphone-quality audio recordings of vehicle acceleration. This was achieved using careful selection of key temporal and spectral features, combined with simple deep learning models. Together, we found schemes that could distinguish between 5 vehicle classes at 75 % accuracy, and 7 vehicle classes at 54 % accuracy. However, the fact that we can achieve high training accuracies using simple models inspires us to close the gap between training and test accuracies. That there is significantly higher training accuracy than test accuracy suggests that we are overfitting to our data in many cases. To mitigate this, we will perform regularization techniques on our models, including drop-out approaches as well as early stopping. Finally, we hope to dive deeper into our RNN architectures to develop those with longer term memory, to capture the inherently time-dependent phenomenon of vehicle acceleration.

7 Github Repository

Please go to `https://github.com/ge0405/CS229_Project.git` to find the codes for this project.

8 Contributions

Both members contributed equally in data acquisition, feature selection, and analysis.

9 References

[1] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. ”Are we ready for autonomous driving? the kitti vision benchmark suite.” 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012.
[2] Odat, Enas, Jeff S. Shamma, and Christian Claudel. ”Vehicle classification and speed estimation using combined passive infrared/ultrasonic sensors.” IEEE transactions on intelligent transportation systems 19.5 (2017): 1593-1606.
[3] Capobianco, Samuele, et al. ”Vehicle Classification Based on Convolutional Networks Applied to FMCW Radar Signals.” Italian Conference for the Traffic Police. Springer, Cham, 2017.
[4] He, Yao, Yuchuan Du, and Lijun Sun. ”Vehicle classification method based on single-point magnetic sensor.” Procedia-Social and Behavioral Sciences 43 (2012): 618-627.
[5] Nooralahiyan, Amir Y., et al. ”A field trial of acoustic signature analysis for vehicle classification.” Transportation Research Part C: Emerging Technologies 5.3-4 (1997): 165-177. [6] George, Jobin, et al. ”Exploring sound signature for vehicle detection and classification using ANN.” International Journal on Soft Computing 4.2 (2013): 29.
[7] Alexandre, Enrique, et al. ”Hybridizing extreme learning machines and genetic algorithms to select acoustic features in vehicle classification applications.” Neurocomputing 152 (2015): 58-68.
[8] Wieczorkowska, Alicja, et al. ”Spectral features for audio based vehicle and engine classification.” Journal of Intelligent Information Systems 50.2 (2018): 265-290.
[9] Libraries: scikit-learn, PyTorch, matplotlib and librosa etc.