An Ensemble Technique for Estimating Vehicle Speed and Gear Position from Acoustic Data

Hendrik Vincent Koops
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
h.v.koops@uu.nl

Franz Franchetti
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
franzf@ece.cmu.edu

***Abstract*—This paper presents a machine learning system that is capable of predicting the speed and gear position of a moving vehicle from the sound it makes. While audio classification is widely used in other research areas such as music information retrieval and bioacoustics, its application to vehicle sounds is rare. Therefore, we investigate predicting the state of a vehicle using audio features in a classification task. We improve the classification results using correlation matrices, calculated from signals correlating with the audio. In an experiment, the sound of a moving vehicle is classified into discretized speed intervals and gear positions. The experiment shows that the system is capable of predicting the vehicle speed and gear position with near-perfect accuracy over 99%. These results show that this system could be a valuable addition to vehicle anomaly detection and safety systems.**

***Index Terms*—Acoustic signal processing; Classification algorithms; Vehicle safety; Automotive applications**

I. Introduction

The sound emitted by a machine is often an indication of the quality and nature of its performance. It can be used to identify the type of operation, failures, or abnormal running conditions. For example, after being exposed to the characteristic sound of their vehicle for a long time, most people are able to distinguish by ear when there is something out of the ordinary, because the sound changed from "normal" to abnormal. Although it might not be immediately apparent what the exact problem is, it can often be used as an indication of an unwanted change. Automating the process of diagnosis by sound can be a very useful addition to currently existing safety systems.

Related work. Automatic audio classification has has been thoroughly studied in different research areas such as music information retrieval (e.g. genre classification), speech recognition and bioacoustics (e.g. birdsong recognition). Most research in the area of audio analysis and machine sounds focusses on automatic failure diagnosis [1]–[3]. Nevertheless, research into predicting the state of a vehicle (such as speed, gear position or engine rpm) from the sound it makes is rare.

Contribution. We propose a novel application of a common audio pattern recognition pipeline that aims to bridge this gap. The machine learning system in this paper is capable of predicting the speed and gear position of a moving vehicle from the sound it makes. Furthermore, we introduce an optimization step to a gradient boosting classifier that in an experiment improves classification to near-perfect results.

Synopsis. The remainder of this paper is structured as follows. Section II introduces the system, and details the three composed steps of which it is created. An experiment using the system can be found in Section III. Section IV will conclude this paper with some final remarks.

II. Methodology

Figure 1 presents the system proposed in this paper in three composed steps: 1) feature extraction, 2) machine learning and 3) optimization. In the first step, features are extracted from portions of audio recordings, which are used as training data for a classification algorithm, together with discretized speed values or gear position information. The machine learning algorithm returns confidence values, or *probabilities* for each possible (speed interval or gear position) class. These probabilities are then used as input for a third step, in which a speed-gear correlation matrix is used to optimize the prediction. The following sections discuss these steps in detail.

A. Feature extraction

The system proposed in this paper aims to classify portions (called *frames*) from audio recordings into one from a set of classes. Extraction of meaningful features to describe the contents of these frames is a critical step towards this goal,

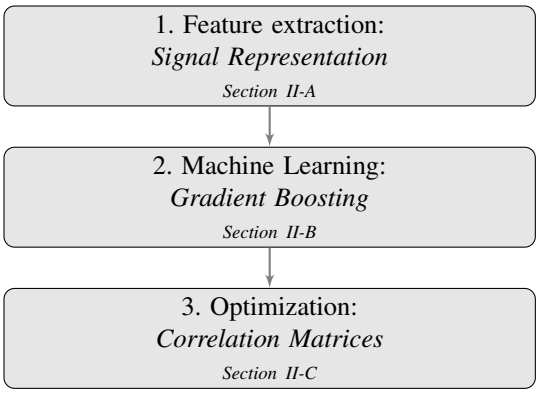

Fig. 1. Pipeline of the system in three composed steps

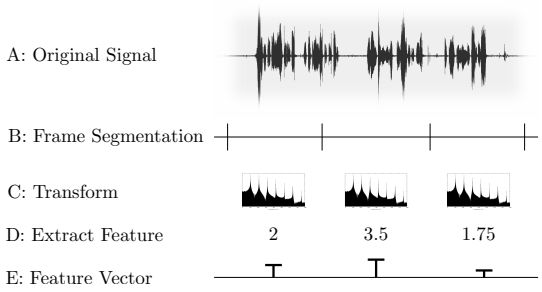

Fig. 2. Example of feature extraction of an audio signal

because direct processing of raw audio data is an enormous task. Feature extraction aims to represent data in such a way that it can be productively used in machine learning tasks by describing the content more efficiently, which in turn relaxes space and time processing requirements. It can be seen as a way to reduce the dimensionality of the data by finding the the most compact and informative set of numerical representations required to to describe a large set of data accurately for the task. Feature extraction is a common step in pattern recognition tasks that involve the processing of sound, and usually involves a number of crucial steps.

A common way to perform feature extraction from audio can be found in Figure 2. Commonly, a signal (A) is broken down into equally sized segments (B). Segmentation of the audio signal allows for creating a representation that varies over time, and the analysis of parts of a signal. Although the figure implies that frames are distinct, they are usually computed in an overlapping fashion, with a new segment starting halfway the size of the frame. Each of the frames is transformed (C), for example by using a Fourier transform, which transforms the frames' representation in the time domain to a representation in the frequency domain. This representation is then described using one dimensional features (D). An example of a simple feature is the average amplitude of a segment, which can be used to distinguish if a frame is loud or quiet. A list of features describing a single frame is called a *feature vector* (E). A combination of feature vectors is called a *feature matrix*.

This feature matrix is used as input to the next part of the pipeline, in which a machine learning algorithm called *Gradient Boosting* is trained to predict the correct class associated with these feature vectors. These classes are values of a discretized signal recorded simultaneously with the audio, such as speed, gear position and engine rpm. The next section describes this machine learning algorithm.

B. Gradient Boosting

To create a class prediction of a feature vector, a machine learning technique called *Gradient Boosting* (GB) is used. This technique was chosen from a large number of classifiers for its ability to outperform other techniques in an initial comparison.

GB is an ensemble machine learning technique introduced by Friedman [4] for regression problems, who later optimized the technique [5]. Like most classification algorithms, GB

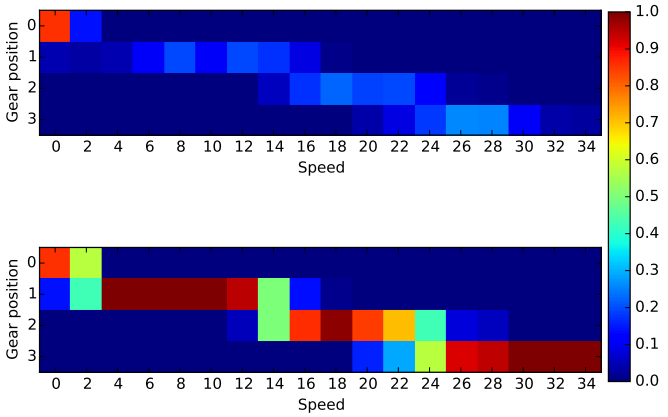

Fig. 3. Example of speed/gear and gear/speed correlation matrices. The x-axis shows the speed at 2 km/h intervals and the y-axis shows the gear position. Low values are blue, high values are red.

tries to find an approximation $\hat{F}(x)$ to a function $F^{\star}(x)$ that connects an output variable y with a vector of input values x. In our case, y is a class such as discretized vehicle speed, and x is a feature vector extracted from an audio frame.

GB constructs additive regression models by stage-wise fitting weak prediction models such as decision trees, thereby improving a non-optimal model F_m at stages $1 \leq m \leq M$ of m iterations. It generalises the models by allowing optimization of an arbitrary differentiable loss function $L(y, F(x))$. In the system proposed in this paper, regression trees are fit on the negative multinomial log-likelihood loss function.

At each stage, F is improved by constructing a new model F_{m+1} that adds an estimator h to construct a better model:

$$F_{m+1}(x) = F_m(x) + h(x) \tag{1}$$

Note that a perfect h implies:

$$F_{m+1}(x) = F_m(x) + h(x) = y \tag{2}$$

Therefore the residual $y - F_m(x)$ is used to fit h, correcting its predecessor.

To train the GB classifier, the set of feature vectors representing the audio signal together with their correlating speed and gear classes are split up in a 70% training and 30% testing set. The classifier is trained on the training set, and the elements of the testing set are used to test the generality of the classifier. The output of GB is a *prediction vector* $CLF = [CLFp_1, CLFp_2, \ldots, CLFp_n]$ for each frame of the audio file, where n is the number of classes used in the training phase and $CLFp_i$ is the classifier probability of class i. Without optimization, the element with the maximal value in the vector would be chosen as the predicted class. In this study, the complete vector is used to improve the classification results. This is done by using a correlation matrix calculated from several classes, which is explained in the next section.

C. Correlation Matrix Optimization

Gear position and speed are correlated, therefore a useful correlation matrix **M** can be calculated that can be used

to improve the classification results. Figure 3 shows two examples of **M**: the probabilities of a gear position coinciding with a discretized speed interval (top) and vice versa (bottom). These example matrices are calculated from 70% of the speed and gear values, in which the speed values are discretized at 2 km/h. The image shows that the probability of being in a higher gear increases with speed, and vice versa. This knowledge can be used to increase the accuracy of the classifier by performing a Hadamard product (point-wise product denoted here as $\odot$) of the classification prediction array with the corresponding row in the matrix.

In the following example, the speed of the vehicle is predicted. First, the gear classifier $CLFg$ is used to return a prediction array $CLFg_k = [CLFg_{k_1}, CLFg_{k_2}, \ldots, CLFg_{k_n}]$ for feature vector k. Next, the trained speed classifier $CLFs$ is used to return a speed prediction array $CLFs_k = [CLFs_{k_1}, CLFs_{k_2}, \ldots, CLFs_{k_m}]$ for a feature vector k. In a next step, the Hadamard product of $CLFs_k$ and $m_{i,*}$ is calculated, where i is the maximum value position in $CLFg_k$, and $m_{i,*}$ is the i^{th} row of $\mathbf{M}_s$. In this case, $\mathbf{M}_s$ is the gear/speed correlation matrix. In this example, row i from Figure 3 (top) will be used in a Hadamard product with $CLFs_k$. This results in a new prediction array $CP = CLFs_k \odot m_{i,*} = [CPp_1, CPp_2, \ldots, CPp_n]$. For gear prediction, a column of the bottom matrix in Figure 3 is used.

D. Evaluation

The system is evaluated in two ways: firstly, we obtain classification results using GB, which is measured as the fraction of correctly classified frames of the audio signal. Secondly, we use using a Hadamard product with a correlation matrix to optimize the GB classification. In this case, a margin of error is introduced with size plus and minus the discretization interval. For example, in the case of a speed discretization interval of d, classification into class i is considered correct if $i - d < c > i + d$, where c is the true class and d is the discretization interval.

III. Results

The following sections describe results obtained in an experiment using the system described in Section II. Section III-A describes the audio recordings and correlating machine states that are used in the experiments. Section III-B discusses results obtained from the GB classifier. After that, results are optimized using correlation matrices (Section III-C).

A. Audio and Vehicle Data

In an experiment to evaluate the system, audio recordings of an American-built car driving multiple identical laps on a closed parking lot are obtained. The vehicle is equipped with a computer, providing real-time recordings of a multitude of states (speed, gear position, rpm, etc). From these machine states the vehicle speed and gear position are used in this experiment. The following two sections describe the audio and vehicle state data.

1) Audio: Three microphones are placed on and inside the vehicle: two piezoelectric microphones and one stereo condenser microphone. The two piezoelectric microphones are attached on the inside and outside of the middle of the windshield, to obtain mechanical vibrations and noise of the car, with and without wind air noise. The stereo condenser microphone is placed in the glovebox to pick up engine noise. In total, four mono audio tracks are obtained. The audio signal is normalized to have mean equal to 0 and maximum absolute value equal to 0.98 to account for differences in amplitude between the four recordings. After normalization, the features are extracted and averaged over the four sources. Twenty five dynamic and spectral features are extracted on 8 different frame sizes: 0.1, 0.5, 1, 1.5, 2, 2.5, 3, and 4 seconds. Some examples of the extracted features are Mel-Frequency Cepstrum Coefficients (MFCCs), zero crossing rate and spectral centroid. MFCCs are often used as features in music information retrieval and speech recognition. They represent the signal on a scale that is closely related to the human auditory systems' response [6]. The zero crossing rate (ZCR) represents the rate at which a signal changes from positive to negative, defined as

$$zcr = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{I}\{s_t s_{t-1} < 0\} \tag{3}$$

where s is a signal of length T and $\mathbb{I}\{\varphi\}$ returns 1 if its argument φ is true and 0 otherwise. It is often used to identify percussive sounds [7].

The spectral centroid represents the center of mass of the spectrum of a signal. It is often associated with the perception of "brightness" of a signal and is defined as:

$$sc = \frac{\sum_{n=0}^{N-1} f(n)\, x(n)}{\sum_{n=0}^{N-1} x(n)} \tag{4}$$

where $x(n)$ represents the weighted frequency value, or magnitude, of bin number n, and $f(n)$ represents the center frequency of that bin. It closely correlates with the perception of brightness of a signal [8]. Discussing the details of all twenty five features is beyond the scope of this paper, see the code repository[1] for a full list.

2) Vehicle State: The driving speed and gear position of the car are recorded simultaneously with the audio. The speed measurements are discretized in intervals of 1, 2 3, 5, 7 and 10 kilometer per hour. Training the system with various granulated data allows for finding an optimal set of parameters. Gear positions are by definition discrete (ranging from 0 to 3) and therefore not altered. The features and the speed and gear vehicle state data are used as input to the GB algorithm in the next step.

B. Gradient Boosting Results

In Tables I and II the results of GB classification results for speed and gear position at different frame sizes are presented.

[1] https://bitbucket.org/hvkoops/machinestateprediction

	0.1	0.5	1	1.5	2	2.5	3	4
10	0.692	0.692	0.706	0.693	0.694	0.572	0.619	0.660
7	0.558	0.583	0.571	0.559	0.573	0.562	0.529	0.521
5	0.529	0.535	0.502	0.530	0.506	0.542	0.481	0.448
3	0.394	0.375	0.360	0.311	0.376	0.385	0.323	0.321
2	0.244	0.233	0.276	0.161	0.245	0.151	0.206	0.139
1	0.156	0.141	0.138	0.118	0.088	0.073	0.079	0.048

TABLE I

SPEED CLASSIFICATION RESULTS USING GRADIENT BOOSTING AT DIFFERENT FRAME SIZES IN SECONDS (COLUMNS) AND SPEED DISCRETIZATION INTERVALS IN KM/H (ROWS)

0.1	0.5	1	1.5	2	2.5	3	4	
10	0.63	0.652	0.652	0.668	0.573	0.61	0.640	0.618
7	0.63	0.652	0.652	0.668	0.573	0.61	0.640	0.618
5	0.63	0.652	0.652	0.668	0.573	0.61	0.640	0.618
3	0.63	0.652	0.652	0.668	0.573	0.61	0.640	0.618
2	0.63	0.652	0.652	0.668	0.573	0.61	0.640	0.618
1	0.63	0.652	0.652	0.668	0.573	0.61	0.640	0.618

TABLE II

GEAR CLASSIFICATION RESULTS USING GRADIENT BOOSTING AT DIFFERENT FRAME SIZES IN SECONDS (COLUMNS) AND SPEED DISCRETIZATION INTERVALS IN KM/H (ROWS)

	0.1	0.5	1	1.5	2	2.5	3	4
10	0.991	0.993	0.992	0.998	0.933	0.945	0.989	0.982
7	0.960	0.961	0.958	0.98	0.873	0.894	0.915	0.939
5	0.888	0.895	0.868	0.914	0.791	0.812	0.836	0.848
3	0.801	0.804	0.762	0.791	0.733	0.754	0.767	0.788
2	0.545	0.560	0.576	0.45	0.506	0.496	0.481	0.485
1	0.347	0.317	0.311	0.264	0.261	0.254	0.217	0.2

TABLE III

SPEED CLASSIFICATION RESULTS USING CORRELATION MATRIX OPTIMIZATION AT DIFFERENT FRAME SIZES IN SECONDS (COLUMNS) AND SPEED DISCRETIZATION INTERVALS IN KM/H (ROWS)

	0.1	0.5	1	1.5	2	2.5	3	4
10	0.976	0.978	0.982	0.989	0.979	0.973	0.968	0.97
7	0.978	0.982	0.977	0.989	0.988	0.973	0.968	0.976
5	0.969	0.971	0.974	0.982	0.873	0.973	0.963	0.927
3	0.967	0.978	0.971	0.98	0.873	0.973	0.963	0.933
2	0.968	0.970	0.980	0.975	0.915	0.973	0.963	0.897
1	0.958	0.965	0.965	0.968	0.945	0.973	0.884	0.854

TABLE IV

GEAR CLASSIFICATION RESULTS USING CORRELATION MATRIX OPTIMIZATION AT DIFFERENT FRAME SIZES IN SECONDS (COLUMNS) AND SPEED DISCRETIZATION INTERVALS IN KM/H (ROWS)

The values represent the fraction of correctly classified audio frames. Note that gear position is not rediscretized, therefore the results are the same for all speed discretizations. The table shows that the speed results improve with increasing discretization step for all the frame sizes. In part, this can be attributed to the decreasing number of classes: with a large discretization step, fewer classes remain.

Overall, the table shows that the speed classification accuracy decreases with an increased frame size and small discretization steps. For larger discretization steps, the results are more stable. The best speed result can be found with a frame size of 1 second and a discretization step of 10, but the difference with similar frame sizes is not significant. The results of predicting the gear position is similar across all frame sizes, and ranges between 0.55 and 0.7. The gear classification is stable for each speed discretization step because the gear positions are not changed.

C. Correlation Matrix Results

Tables III and IV show the results of optimizing the GB results using correlation matrices, as described in Section II-C. The images shows the results improve dramatically for both speed and gear prediction compared to the GB results. Overall, for speed, the same trend is observed: the results improve with an increase of the discretization step. Just as the GB results, the accuracy decreases with an increasing frame size, although the difference is small.

The best speed results can be found with a discretization step of 10, with minimal difference with a discretization step of 7. The optimal value in terms of minimal discretization step and frame size is 0.914, where the discretization step is 5 and the frame size is 1.5. An example of these predicted with error margin can be found in Figure 4. The gear prediction results are again more stable across frame size and speed discretization step, with the best results appearing around a frame size of 1.5.

IV. CONCLUSIONS

This paper proposed a simple yet effective system that is capable of classifying the speed and gear position of a driving vehicle from the sound it makes. After feature extraction, the system uses a 2-step approach of Gradient Boosting and optimization using correlation matrices to predict the speed of the vehicle. In an experiment using sound from a real car, it was shown that near perfect (over 99%) classification results are obtained when the speed is discretized between 10 and 5 km/h, and very high (over 90%) classification results are obtained when speed is discretized on smaller intervals. The classification results show that this system woul be a valuable addition to anomaly detection and safety systems for vehicles.

ACKNOWLEDGMENTS

This work was sponsored by DARPA under agreement HR0011-13-2-0007. The content, views and conclusions presented in this document do not necessarily reflect the position or the policy of DARPA. The authors would like to warmly thank Darrel J. Van Buer, Gavin Holland and Aleksey Nogin at HRL Laboratories for their help with data collection and use of equipment.

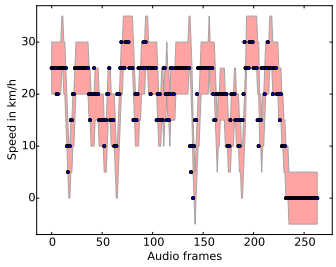

Fig. 4. Example of predicted speeds (blue) agains the actual speeds with error margin (red). The y-axis shows the speed in 5 km/h intervals and the x-axis shows the time expressed in audio frames.

REFERENCES

[1] Z. Li, S. Akishita, and T. Kato, "Engine failure diagnosis with sound signal using wavelet transform," tech. rep., SAE Technical Paper 970034, 1997.

[2] M. Madain, A. Al-Mosaiden, and M. Al-khassaweneh, "Fault diagnosis in vehicle engines using sound recognition techniques," in *Proc. IEEE Int. Conf. Electro/Information Technology (EIT), Illinois State University, Normal, IL, USA*, pp. 20–22, 2010.

[3] J.-D. Wu, E.-C. Chang, S.-Y. Liao, J.-M. Kuo, and C.-K. Huang, "Faults classification of a scooter engine platform using wavelet transform and artificial neural network," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, pp. 18–20, Citeseer, 2009.

[4] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.

[5] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[6] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[7] F. Gouyon, F. Pachet, O. Delerue, *et al.*, "On the use of zero-crossing rate for an application of classification of percussive sounds," in *Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy*, 2000.

[8] G. Peeters, "A large set of audio features for sound description (similarity and description) in the cuidado project," *IRCAM, Paris, France*, 2004.