"Quiet down!": A study on psychoacoustic annoyance within vehicles

Ismael Villegas[1], Erickson R. Nascimento, Ph.D.[2], and Ruzena Bajcsy, Ph.D.[3]

***Abstract*— Much research has gone towards autonomous driving during the past few years. However as we continually increase our focus on the vehicle, we need to remember why were doing this in the first place: the people inside the car. Seeing as that a world with everybody being driven by a fully autonomous car is most likely very, very far into the future, we are focusing on semi-autonomous vehicles and how they can help make changes to the environment based on the human perception of vision and sound. Although the autonomous vehicle research community has been focusing primarily with computer vision, we decided to combine this with psychoacoustics - specifically annoyance and how it affects the drivers focus. The sound in vehicles is not only informative of the state of the vehicle and the environment, but it can also affect a drivers attention, performance, and pleasantness of driving. To our understanding, this might be the first dataset that combines and focuses on not only the visual aspect of driving, but also on the essential auditory component that comes with this experience. Our goal is to create an intelligent agent that acts to improve the drivers pleasantness through acoustic-driven learning. The Berkeley Audio-Visual Dataset is a novel inner and outer vehicle image and sound dataset containing over 3 hours of driving footage, of which over 100K images inside the vehicle have been annotated. Alongside this data, we have gathered a diverse collection of data including 3D point clouds, vehicle's inertial measurements, and vehicle information such as wheel speeds, gear used, among other information in a drive.**

***Index Terms*— psychoacoustic metrics, acoustic-driven, deep reinforcement learning, safety, pleasantness, annoyance**

I. INTRODUCTION

First and foremost, the main system that that is required for driving is safety. Without a safety system to predict and respond to various different environment states and mixed behaviors of human agents, it is unlikely that any person would want to get into the vehicle. The research community has been continuously working towards this goal since the 1920's [1] with relatively great success in modern day autonomous vehicles. While we must take the safety of the vehicle into account, it is also important to consider the overall driving experience in regards to comfort and pleasantness. As the research community works towards the creation of an autonomous vehicle, it is imperative to note that this intelligent system needs to improve the overall in-car experience for the driver. There is no point in creating such a vehicle if nobody would want to use it.

One such factor of improving the in-car experience for a driver is the vehicle's interior sound. This aspect will provide the state of the overall environment and can affect the physiological behavior of drivers. Not only will this help us understand the vehicle's state and environment, but this can affect a driver's performance, attention, and comfort while driving.

Sounds can create various different effects on a human. Just as it can create a pleasant environment and driving experience, it can also help create a stressful situation for the drive which can ultimately increase the probability of traffic accidents due to the physiological and psychological effects of sounds on humans. A study by Fagerlönn [2] investigated the influence of urgent alarms on truck drivers. In this study, they found that drivers would brake significantly harder with a high-urgency warning as compared to a low-urgency warning. Ho and Spance [3] found that a simple auditory signal such as a 2 kHz tone is itself capable of distracting a driver. While these studies were done specifically for driving, Sammler et al. [4] presented a study inspecting the electroencephalogram (EEG) power and heart rate change with pleasant and unpleasant states induced by consonant and dissonant music. In music theory the quality of consonance is described as pleasant and agreeable, whereas dissonance is seen to be harsh. It is described to cause tension and desire to be resolved to a consonant interval. This is not an exhaustive list for the neural bases of emotion and mood, and the psychology of pleasantness. If interested we refer the reader to the works of Ruckmick [5] and Dalgleish [6].

Over the past two decades we have seen an increasing body of research of the sound quality of vehicle interior noise. Thus, the consumer has become highly sensitive to a variety of vehicle sounds such as engine sound, warning chimes, door sounds, radio, etc. As of now, the vehicle's sound characteristics is one of the most relevant factors affecting customer vehicle preferences [7]. Past works in vehicle sound quality include various facts, such as quietness and sound pleasantness, but these works have been mainly focused on certain requirements for the design and production of new vehicles. Research also shows that there is a complex tradeoff between removing disturbing sounds and the expectations of the consumer regarding the sound quality of a specific brand and/or model of the vehicle. While there are times where quieter is better [8], quietness is not always the ultimate goal. Quietness is seen as undesirable most of the time to avoid creating a monotonous environment inside the vehicle [7]. In fact, there are products sold in order to

[1]I. Villegas is an undergraduate student of computer science at Columbia University, New York, NY 10027, USA iv2181 [at] columbia [dot] edu

[2]E. R. Nascimento is with the Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG 31270, Brazil erickson [at] dcc [dot] ufmg [dot] br

[3]R. Bajcsy is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA bajcsy [at] eecs [dot] berkeley [dot] edu

increase the sound of a car and break through the silence.

Seeing as that sound plays a large factor to the driving experience, this begs the question of where a large-scale dataset exists specifically for visual *and* audio data inside a vehicle. There has been a recent revolution in computer vision due to the success of deep learning [9], but in order for us to properly train these models we need a significant amount of annotated data and computational resources. We have seen an expanding amount of these large-scale datasets regarding the visual aspect of a drive such as *CamVid* [10], *KITTI Vision Benchmark Suite* [11], *Cityscapes* [12], *Leuven* [13], *Daimler Urban Segmentation* [14], etc. However, of the aforementioned datasets none take focus on the driver or the inside environment. Other projects have collected data looking towards the driver such as Brains4Cars [15] and MIT-AVT [16], but these also follow the trend of not making acoustics an essential portion of their data.

Another recent trend has been integrating synthetically-rendered data from sources like the *Grand Theft Auto V engine* [17] which we began using before having access to physical vehicles. We previously had used the Grand Theft Auto (GTA) V (Rockstar Games, NY) game to generate different interactive scenarios. During the task, the agent was exposed to various sounds due to environmental factors and other external auditory stimuli, such as pedestrians and traffic. We also took control of the radio and included a passenger who would speak on command to create a conversational environment inside the vehicle. Soon afterwards, we were offered the use of one of the cars, the Lincoln MKS, at Berkeley DeepDrive and promptly started collecting real-life data.

As a result, in this paper we present an ongoing, novel, audio-visual project aimed towards providing a diverse and comprehensive dataset for a vehicle's interior and exterior. Our dataset has the following characteristics:

1) This first subset is over 100K image frames with annotation. We divided our dataset into different experimental drives between different days. This was in order to keep similar weather conditions grouped together.

2) Our dataset has survey-grade dense 3D point cloud for static objects as well as inertial measurements, wheel speeds, and a variety of the vehicle's information.

3) Annotated events happening inside the vehicle (such as people speaking, radio state, etc) as well as bounding boxes over anybody inside the vehicle.

II. Related Work

As mentioned before, the safety of drivers has improved over decades of research through driver modeling studies [19 - 21] that develop the understanding of predicting driver's behaviors. Despite these advances in visual perception and driver modeling, there still exists an absence of an essential component to the vibro-acoustical system: the driver, passengers, and vehicle. This work builds upon findings from sound recognition and psychoacoustic sound evaluations. A driver's emotional state has been recognized to significantly affect the safety of driving.

Most works on noise treatment inside the vehicle have focused solely on noise detection or reduction for adjustments during the design of the vehicle. Several psychoacoustic indices have been introduced in sound quality evaluation engineering [22], [23] and evaluated in various driving scenarios. For instance, in Nor et al.'s [24] studies both subjective and objective tests are conducted to evaluate vehicle comfort index. They found that the metrics of loudness, sharpness, roughness, and fluctuation strength are correlated with human subject studies. Additionally, they showed how the acoustical comfort is affected by each of these metrics.

Another notable study on noise in a vehicle's interior was Duan et al. [25] who studied the acoustics in different conditions of the vehicle: idle, constant speed, accelerating, and braking. They proposed to predict the sound quality by using a neural network. The input features used were four psychoacoustic parameters: loudness, sharpness, articulation index, and A-weighted sound pressure levels (SPLs) and subjective annoyance was used as a label to train the network. They then created a dataset composed of 36 interior noise samples from the vehicle in the aforementioned vehicle conditions. The measurement followed the GB/T 18697 standard and their results showed an accuracy above 95.57%.

Although there have been several psychoacoustic indices proposed in the past years, when one is driving this information alone is not useful for identifying the sound source of annoyance. A candid way of going about this problem is to separate the sound events and compute the annoyance of each one, but the detection and segmentation of general sound events is still not a well-solved problem.

A challenging task in sound recognition is the classification of individual sounds in composite sounds. The task is known as polyphonic sound event detection and consists of detecting multiple overlapping sound events. There have been several approaches proposed including transfer learning [26], cross-modal learning [27], and deep learning methods [28].

One particular problem of learning from sound samples is the lack of large-scale annotated data. This type of absence is particularly true for sounds from the vehicle interior as most of the available sound datasets are composed of ambient, event, or mixed sounds. Apart of being influenced by various uncorrelated and dynamic factors, the sound in the interior of the vehicle also affects people differently. Thus, we focus on creating a dataset to test our learning algorithms.

III. Methodology

In order to collect this data, we used the vehicle in Figure 3 courtesy of Berkeley Deep Drive. The specifications of the vehicle and hardware used to collect data can be found in the section labeled "Hardware." All drives were conducted by the investigators and each different event or internal environment change is separated by an audible clap from one of the

(a) Bounding Boxes (b) Heat Map

Fig. 1: Annotated Frames for each frame of video. Bounding boxes are given to localize where a sound will be coming from and a heat map is done to localize sound and annoyance.

investigators in order to make synchronization simpler for those using the data.

We used various different routes inside the Richmond Field Station in Richmond, CA to get the environment of a quiet park as well as paved roads and dirt roads. From this station, we drove to University of California, Berkeley campus in order to capture data on the freeway and in downtown Berkeley. These different environments allow us to gather various different outside scenarios whether this be honks from other cars, ambient speech from pedestrians passing by the vehicle, or the sounds of afternoon traffic.

The initial focus that we investigated were the auditory aspects caused by actions that could be taken by an agent. That is, we first decided to analyze controlled environments that an autonomous car could control. We created different scenarios that included various combinations of states from the windows, radio, air conditioner, and speeds. During most of these experiments, we decided to stay silent inside the car so that we have data where our agent could focus on the aspects it could control.

It is also imperative to treat the driver and passengers as components of the vibro-acoustical system since they also contribute to the overall soundscape inside the vehicle. Therefore, in order to collect a variety of soundscapes we created scripted events for the passengers and driver of the vehicle. For example, in one drive a driver might be asked to talk amongst themselves for two minutes while all others inside the car remained quiet. This could be done for all other passengers. We also recorded instances of conversations between two different people and instances of people speaking at the same time. Annotations of where a sound is coming from and who is speaking is shown on Figure 1.

Apart from collecting our data, we implemented an exploration-exploitation approach based on deep reinforcement learning where an agent learns actions to decrease the annoyance – further improving an existing framework to run on our simulations and collected data. Seeing as that the sound in the vehicle's interior is influenced by several

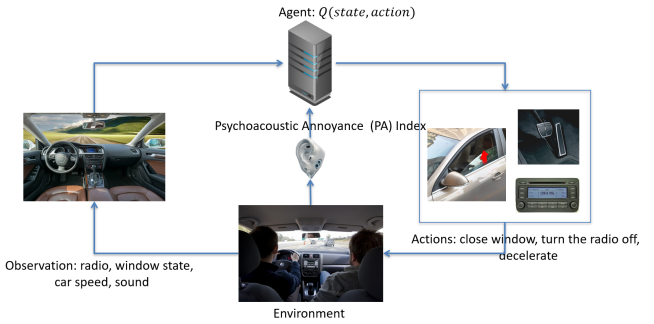

Fig. 2: Our approach is based on a deep reinforcement that learns from the environment how to change the state of the vehicle to avoid an unpleasant environment. After the agent took action, a_i, considering the environment state, s_i, our system computes the reward, r_i, using the psychoacoustic annoyance (PA) metric. The reward, the action, and the environment states are used to train a neural network that approximates the Q-function. The neural network is posteriorly used to decide which action the agent should take to decrease annoying sound.

uncorrelated and dynamic factors, our agent can learn from the environment how to change the state inside the car to reduce the psychoacoustic annoyance levels for the driver. Figure 2 depicts the main steps of the learning method.

A. Learning

In our paper, we formulate the problem of choosing an action that avoids annoying sounds as a Markov Decision Process (MDP). Let $A = a_1, \ldots, a_n$ be a discrete action set and S the state set where the agent takes action a_i considering the environment state s_i represented at the i-step of the episode.

An action a_i leads to a state transition from the state s_i to s_{i+1} and an immediate reward r_i. In order to maximize this accumulated reward R defined as

$$R = \sum_i \gamma^{k-1} r_i, \tag{1}$$

where $\gamma \in [0, 1]$ is the discount factor for future rewards, the learning process tries to minimize the loss function:

$$L = \frac{1}{2}(r + \max_{a'} Q(s', a') - Q(s, a))^2, \tag{2}$$

where $Q(s, a)$ is the function that gives the best score after performing the action a on the state s.

Our approach used a ϵ-greedy strategy to explore the action space and learn the result of each action on the annoyance metric. With a probability ϵ, we randomly select an action and with probability $1-\epsilon$ we follow the action that maximizes the quality of current and future actions. Thus, probability $1 - \epsilon$ the action is determined by the policy

$$\pi(s_i) = \arg\max_a Q(s_i, a). \tag{3}$$

The Q-function is modeled by a multilayer perceptron network, which receives as input a state vector and returns a vector containing the Q-value for each possible action. This network is initialized with random parameters.

B. Immediate Reward

The sound impression for a human listener can be estimated by the following psychoacoustic properties [29]:

- **Fluctuation** and **Roughness**: When we have multiple signals with different frequencies in an environment, they might interfere constructively and destructively with each other creating modulation. In other words, the amplitude of a sound signal rise and fall over time. Fluctuation and roughness measure the modulation of signal over time. Fluctuation, F, was designed to work with up to 20 modulations per second and can be given by:

  $$F \approx \frac{\Delta L}{4Hz/f_{mod} + f_{mod}/4Hz}, \tag{4}$$

  where ΔL is the modulation depth and f_{mod} the modulation frequency. The roughness R describes sounds with modulations range from 20 to 300 times per second, and can be computed as being the product:

  $$R \approx \Delta L \times f_{mod}. \tag{5}$$

  A modulated signal is more unpleasant when having a higher roughness and fluctuation;
- **Loudness**: The loudness is not a physical phenomenon, rather a psychological phenomenon which is based on perceived loudness. Differently from the sound level that is a physical measurement, the loudness was developed based on human subject studies in persons with normal hearing. Each person listened to a tone at frequency f and a particular dB level, a second tone would be played at a different frequency. The level of the second tone would be altered until it sounded equally as loud as the f tone. Let E_{TQ} be the excitation at threshold in quiet, and E_0 e the excitation of the reference intensity, the specific loudness of a sound with excitation E is given by

  $$N' = 0.08(\frac{E_{TQ}}{E_0})[(0.5 + 0.5\frac{E}{E_{TQ}}) - 1]. \tag{6}$$

  The total loudness is the result of integrating the specific loudness over critical-band, the smallest band of frequencies that activates the same part of the basilar membrane in the human hearing system, rates, i.e.

  $$N = \int_0^{24} N' dz, \tag{7}$$

  where z is the critical-band in Bark.
- **Sharpness**: A function of the spectral composition. It is estimated by a weighted sum of specific loudness levels in different bands. The total sharpness is given by

  $$S = \int_0^{24} S' dz, \tag{8}$$

  where $S' = \frac{0.11}{N} \int_0^{24} N' g(z) z dz$ is the specific sharpness and $g(\dot{})$ is a critical-band-rate dependent weighting function. The sound with higher sharpness is more unpleasant and annoying.

The sound annoyance is closely related to the aforementioned psychoacoustic indices. Zwicker proposes to compute the psychoacoustic annoyance (PA) [29] as a function of sharpness, loudness, fluctuation, and roughness as follows:

$$PA = N_5(1 + \sqrt{\omega_S^2 + \omega_{FS}^2}), \tag{9}$$

where N_5 is the 95th percentile of loudness and

$$\omega_S = \begin{cases} -(S - 1.75)log(N_5 + 10), & S > 0 \\ 0, & otherwise, \end{cases} \tag{10}$$

$$\omega_{FS} = \frac{2.78}{N_5^{0.4}}(0.4F + 0.6R). \tag{11}$$

We define the immediate reward as being a function of psychoacoustic annoyance metric, i.e., $r_i = f(PA)$ is given by the PA metric, where $f(\dot{})$ is the shape function $f(x) = 1 - (\frac{x}{MAX_{PA}})^{0.4}$ and MAX_{PA} is the maximum acceptable value for PA. In our experiments we used $MAX_{PA} = 27$.

IV. HARDWARE

The most essential portion to collecting data for naturalistic driving is having the appropriate hardware and software to efficiently grab the information of your surroundings. We were fortunate enough to use one of the vehicles, a Lincoln MKS, owned by the Berkeley Deep Drive Team shown in Figure 3.

This vehicle was designed to satisfy the following goals and requirements:

1) **Timestamped Sensor Recording**: Recording all mounted sensors and data streams in a way that each sample of data is timestamped using the centralized and reliable timekeeper in the Robotics Operating System (ROS). The resulting data is saved into a .ros file to work with all the captured data in one central location.

2) **High Resolution Video**: Capture and record one to eight mounted cameras at 720p (2.1 megapixels) resolution at 60 frames per second (fps). The camera

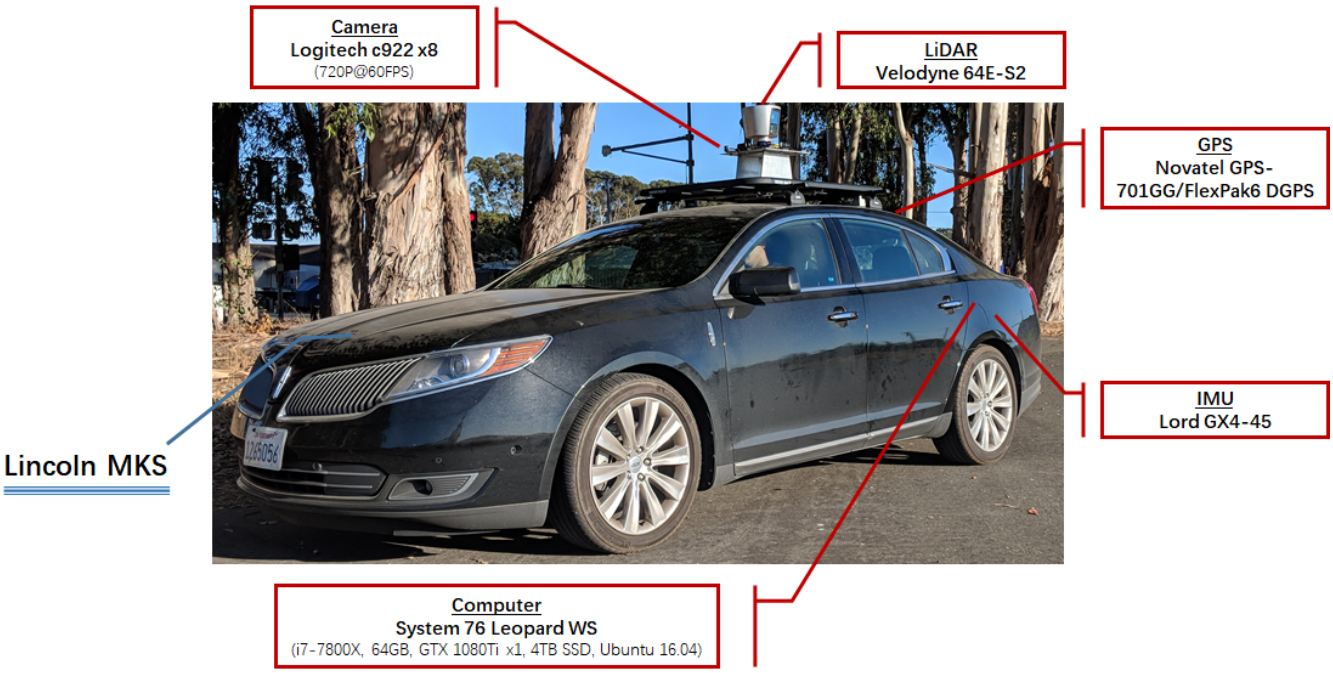

Fig. 3: The vehicle provided by Berkeley Deep Drive. An annotated picture for visualization of where each instrument lies on the vehicle.

position, resolution, and compression were already set when we were given access to the vehicle.

3) **CAN Bus**: Collect vehicle information from the Controller Area Network (CAN) bus of the vehicle [18]. Every vehicle has different ports and bus utilization policies. These raw CAN messages are recorded in various different ROS topics for organization.

4) **Nonintrusive Design**: The system is designed in such a way so that the driver experiences as close to a naturalistic drive as possible. In order to capture realistic data the design was done so that the driver does not have to modify their driving for the vehicle.

The platform is equipped various tools to meet the aforementioned requirements. Figure 3 has been provided for the visualization of where each instrument may lie. The instruments used are as follows:

- **Computer**: A computer was installed in the trunk of the vehicle to avoid driver disruptions and control the other instruments used. The computer is a System76 Leopard WS with an Intel i7-7800X processor, 64 GB DDR4, 4 TB SSD, and an Nvidia GeForce 1080Ti.
- **Cameras**: Mounted atop the vehicle are eight Logitech C922 video cameras pointing outwards towards the street in a circular formation to get a 360° view of the vehicle.
- **GPS**: A Novatel GPS-701-GG/FlexPak6 DGPS is used to record the GPS information.
- **LiDAR**: A Velodyne 64E-S2 was used as our surveying tool to get the depth of our surroundings.
- **IMU**: A Lord GX4-45 was used as our Inertial Measurement Unit (IMU).

Unfortunately, those who created the vehicle specifications did not take auditory signals or inner car information into account when building the instrument. Therefore, we provided our own instruments to take the necessary data for the dataset. We utilized a GoPro Hero 5 and two Olympus ME-51S stereo microphones with deadcats for each microphone to reduce wind noise. One of the microphones was connected directly to the GoPro in order to synchronize the inside noise with the inside video. The other microphone was connected to the on-board computer to record sounds happening outside of the vehicle.

V. Trips and Files

This section will define how trip data files may be stored in a trip directory. A trip directory means that a driver took their vehicle from start to finish. These are files that are extracted from either a GoPro Hero or the Lincoln MKS computer and placed into a central server to be filtered, cleaned, synchronized, and processed.

A. Trip Day Folders

Trip day folders are directories that separate drives by any given day. This was done so as to have an organized platform that can be separated by any environmental conditions that cannot be controlled (sunny, overcast, etc). All will be formatted in the following manner:

- **YYYY-MM-DD**: These directories will hold all the experiments done on the day of the directory name.

B. Trip Data Files

Trip data files are the endpoint of all streams from every sensor for each drive. This includes many CSV (comma separated values) with timestamped information. The trip data files are separated between file types for consistency and organization as follows:

- **Rosbag**: These .ros files contain a variety of information from any given car ride. They are as follows:
  - **velodyne_packets**: Raw data packets from Velodyne LiDAR sensor. Captured at 10 Hz.
  - **velodyne_points**: Accumulated Velodyne points transformed in the original frame of reference. Includes fields for "intensity" and "ring." Captured at 10 Hz.
  - **nmea_sentence**: GPS sentences with the data type GGA. This includes the essential current fix data which provide 3D location and accuracy. Captured at 182 Hz.
  - **usb_camX/camera_info**: Includes overall camera information for all 8 usb cameras. Captured at 30 Hz.
  - **usb_camX/image_raw/compressed**: Compressed raw images for all 8 usb cameras. Captured at 30 Hz.
  - **vehicle/acc_ped_eng**: Includes throttle_rate, throttle_pc, and engine_rpm (revolutions per minute). Captured at 101 Hz.
  - **vehicle/brake_ped**: Shows if brake pedal is being pressed. Captured at 50 Hz.
  - **vehicle/brake_torq**: Includes brake_torque_request, brake_torque_actual, and vehicle_speed. Captured at 50 Hz.
  - **vehicle/gear**: Which gear the vehicle is currently using. Captured at 50Hz.
  - **vehicle/imu/data_raw**: The orientation, angular_velocity, and linear_acceleration along with the covariance for each of the three aspects. Captured at 50 Hz.
  - **vehicle/joint_states**: Position and velocity of each wheel and the steer on the front left and front right. Captured at 116 Hz.
  - **vehicle/steering_ang**: Steering wheel angle. Captured at 50 Hz.
  - **vehicle/steering_torq**: Steering wheel torque. Captured at 50 Hz.
  - **vehicle/suspension**: Front and rear suspension. Captured at 50 Hz.
  - **vehicle/tire_press**: Tire pressure for each wheel. Captured at 2 Hz.
  - **vehicle/turn_sig**: Whether turn signal was being used or not. Captured at 1 Hz.
  - **vehicle/twist**: Linear and angular twist of entire vehicle. Captured at 50 Hz.
  - **vehicle/wheel_speeds**: Records the speed of each of the four wheels in miles per hour. Captured at 100 Hz.
- **CSVs**: The .csv files are extracted information from the .ros files with the timestamp (seconds and nanoseconds) for readable format without the necessity of using ROS. As of now the only information extracted is nmea_sentences, vehicle/gear, and vehicle/wheel_speeds.
- **GoPro**: Includes the original .mp4 files from each drive along with an .lrv (low resolution video) version.
- **Processed Video**: Full, uninterrupted .mp4 video from GoPro since they are automatically split after approximately 15 minutes.
- **Images**: The .jpg files which correspond to the frames of the .mp4 files at 10 frames per second.

C. Filtering Criteria

The main focus of this project was to grab a large amount of diverse data, but our main priority was to capture audio and visual data inside the vehicle. Periodically, and for some unexplained reason, the GoPro would record the visual portion of the video as expected, but the audio was not recorded whatsoever. Therefore, any drive taken that had the audio dropped was not included in our dataset. The video of the inside of the vehicle and its remaining data has been retained in the event that anybody would need such data.

VI. CONCLUSIONS

Our overall contribution from this study is a diverse, comprehensive, novel driving dataset with annotations. This dataset comes with comprehensive annotations that are necessary for a complete driving system. Alongside the data on the interior of the vehicle, we collected a variety of the vehicle's information spanning from the wheel speeds to the depth of its surroundings.

ACKNOWLEDGMENT

We would like to thank Frank Zheng for his contributions to this project in the lab. We would also like to thank Berkeley Deep Drive for their contribution of the vehicle used for experimentation.

REFERENCES

[1] K. Bimbraw, "Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology," 2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO), Colmar, pp. 191-198, 2015.

[2] J. Fagerlönn, Urgent alarms in trucks: effects on annoyance and subsequent driving performance, IET Intelligent Transport Systems, vol. 5, no. 4, pp. 252258, 2011.

[3] C. Ho and C. Spence, Assessing the effectiveness of various auditory cues in capturing a drivers visual attention. Journal of experimental psychology: Applied, vol. 11, no. 3, p. 157, 2005.

[4] D. Sammler, M. Grigutsch, T. Fritz, and S. Koelsch, Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music, Psychophysiology, vol. 44, no. 2, pp. 293304, 2007.

[5] C. A. Ruckmick, The psychology of pleasantness. Psychological Review, vol. 32, no. 5, p. 362, 1925.

[6] T. Dalgleish, The emotional brain, Nature Reviews Neuroscience, vol. 5, no. 7, p. 583, 2004.

[7] K. Bijsterveld, E. Cleophas, S. Krebs, and G. Mom, "Sound and safe: a history of listening behind the wheel." Oxford University Press, 2014.

[8] Krebs, S., "The French Quest for the Silent Car Body: Technology, Comfort and Distinction in the Interwar Period." Transfers. Interdisciplinary Journal of Mobility Studies 1, 3, 6489, 2001.

[9] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning." Nature, 2015.

[10] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. "Segmentation and recognition using structure from motion point clouds." In *ECCV*, pages 4457. 2008.

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. "Vision meets robotics: The KITTI dataset." *IJRR*, 2013.

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. "The Cityscapes dataset for semantic urban scene understanding." In *CVPR*, 2016.

[13] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. "Dynamic 3D scene analysis from a moving vehicle." In *CVPR*, 2007.

[14] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth. "Efficient multi-cue scene segmentation." In *GCPR*, 2013.

[15] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, A. Saxena. "Car that Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models." In *ICCV*, 2015.

[16] L. Fridman, D. E. Brown, M. Glazer, W. Angell, S. Dodd, B. Jenik, J. Terwilliger, J. Kindelsberger, L. Ding, S. Seaman, H. Abraham, A. Mehler, A. Sipperley, A. Pettinato, B. Seppelt, L. Angell, B. Mehler, B. Reimer. "MIT Autonomous Vehicle Technology Study: Large-Scale Deep Learning Based Analysis of Driver Behavior and Interaction with Automation."

[17] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. "Playing for data: Ground truth from computer games." In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, (ECCV), volume 9906 of LNCS, pages 102118. Springer International Publishing, 2016.

[18] R. Li, C. Liu, and F. Luo, A design for automotive can bus monitoring system, in Vehicle Power and Propulsion Conference, 2008. VPPC08. IEEE. IEEE, 2008, pp. 15.

[19] C. Little, The intelligent vehicle initiative: advancing human-centered smart vehicles, Public Roads, vol. 61, no. 2, pp. 1825, 1997.

[20] E. Wahlstrom, O. Masoud, and N. Papanikolopoulos, Vision-based methods for driver monitoring, in Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE, vol. 2. IEEE, 2003, pp. 903908.

[21] K. Driggs-Campbell, V. Govindarajan, and R. Bajcsy, Integrating intuitive driver models in autonomous planning for interactive maneuvers, IEEE Transactions on Intelligent Transportation Systems, vol.18, no.12, pp. 34613472, 2017.

[22] H. Fastl, The psychoacoustics of sound-quality evaluation, Acta Acustica united with Acustica, vol. 83, no. 5, pp. 754764, 1997.

[23] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in Proc. IEEE ICASSP 2017, New Orleans, LA, 2017.

[24] M. J. M. Nor, M. H. Fouladi, H. Nahvi, and A. K. Arifn, Index for vehicle acoustical comfort inside a passenger car, Applied Acoustics, vol. 69, no. 4, pp. 343353, 2008.

[25] Z. Duan, Y. Wang, and Y. Xing, Sound quality prediction of vehicle interior noise under multiple working conditions using back-propagation neural network model, Journal of transportation Technologies, vol. 5, no. 02, p. 134, 2015.

[26] W. Han, E. Coutinho, H. Ruan, H. Li, B. Schuller, X. Yu, and X. Zhu, Semi-supervised active learning for sound classication in hybrid learning environments, PloS one, vol. 11, no. 9, p. e0162075, 2016.

[27] D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., SoundNet: Learning Sound Representations from Unlabeled Video. Curran Associates, Inc., 2016.

[28] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, Convolutional recurrent neural networks for polyphonic sound event detection, IEEE/ACM Trans. Audio, Speech & Language Processing, vol. 25, no. 6, pp. 12911303, 2017.

[29] E. Zwicker and H. Fastl, Psychoacoustics: Facts and models. Springer Science & Business Media, 2013, vol. 22