Acoustic-driven Interior Vehicle Adaptation based on Deep Reinforcement Learning to Improve Driver's Comfort

Anonymous submission

***Abstract*—The safety and comfort of drivers have been improved over the decades through new technologies and driver modeling studies that broadened understanding of predicting driver's behaviors. Despite remarkable advances in autonomous systems and interactive systems, there is a significantly lack of approaches considering the passengers and the vehicle as components of a dynamical vibro-acoustical system. The sound in vehicles is not only informative of the state of the vehicle and the environment, but it can also affect driver's performance, attention, and pleasantness of driving. This project aims to investigate the interplay between the perceived sounds of a vehicle and the psychoacoustic annoyance metrics. Our goal is to create an intelligent agent that acts to improve the driver's pleasantness through acoustic-driven learning. To tackle the problem of acting to reduce the annoyance, we present a method based on reinforcement learning that learns from the environment, i.e., the vehicle interior. Our method changes the state inside the vehicle (closing or opening the window and choosing the cruise speed) to avoid annoying sounds in its interior. The results of this work, all performed using the GTA V simulator, showed that the trained agent learned to take actions to avoid creating annoying sounds.**

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

I. Introduction

A challenge in developing safety systems for driving is predicting and responding to the environment state and the behaviors of human agents. These systems must interact in mixed environments and be aware of different sources of information about the environment state such as acoustical and visual signals. In addition to ensuring safe and effective autonomous systems, it is also essential to take into account the pleasantness and comfort of the driving experience. While there is a great deal of work going into bringing autonomous vehicles into fruition, it is noteworthy that an intelligent assistance system should focus on improving the in-car experience for the driver, which is one of the most critical factors on the complex transportation system.

In this paper, we go beyond safety systems and by learning how to accommodate the driver best. An important factor to consider is the vehicle interior sound, which provides information about the state of the environment and can impact the physiological behavior of drivers. The sound in vehicles is not only informative of the state of the vehicle and the environment, but it can also affect driver's performance, attention, and comfort of driving. After all, when driving, we are secluded from the surrounding world by the vehicle, which decreases our sensory experience of the world and increases the risk of accidents.

While on one hand sound can provide a calming effect on the driver and improve their driving experience, on the other hand, the vehicle interior noise may increase the probability of traffic accidents because of the physiological and psychological effects of sounds on the human. Fagerlönn [1], for instance, investigated the influence of urgent alarms in truck drivers. The authors reported that when the drivers received a high-urgency warning, they braked significantly harder when compared to a low-urgency warning. Ho and Spance [2], for their turn, showed that simple auditory signal such as a 2 kHz tone is capable of distracting drivers. Sammler et al. [3] presented a study on electroencephalogram (EEG) power and heart rate change with pleasant and unpleasant states induced by consonant and dissonant music. While the consonance quality is associated with sweetness and pleasantness; dissonance is associated with harshness and unpleasantness. In the context of this paper, *pleasantness* concerns a conscious state related to the satisfaction of a person while driving a vehicle. For a complete picture of the neural bases of emotion and mood, and the psychology of pleasantness, refer the reader to the works of Dalgleish [4] and Ruckmick [5].

The past two decades have witnessed a growing body of research on the sound quality of vehicle interior noise. The customers thus became highly sensitive to sounds relating to vehicles, such as engine sound, warning chimes, door sounds, vehicle audio system, etc. The vehicle sound characteristics are one of the most relevant factors affecting customer vehicle model preference [6]. Previous works in vehicle sound characterization include various aspects, from quietness to sound pleasantness, but have primarily been applied to set requirements for the design and production of new vehicles. However, there is a complex tradeoff between the elimination of disturbing noises and the expectations of the listener concerning the sound quality, of a specific brand and model of the car. Quietness is not always the goal, being most of the time undesirable to avoid creating a monotonous environment in car's interior [6].

In a joint effort, a consortium of five companies and two universities partnered in the OBELICS project [7] to create a sound dataset covering 14 driving condition and 15 vehicles aimed to understand the sound language and sound perception of drivers. The focus of the OBELICS project was the interior noise of vehicles. Several tools were developed in the OBELICS project, such as a tool for extracting and recomposing harmonic and non-harmonic components of sounds and a psychoacoustic parametric synthesizer.

Although the human perception of sound as pleasant or annoying is very subjective, some general findings in human subject studies have resulted in the determination of several psychoacoustic properties. Due to individual differences and context of sound exposure, different people may have a different perception of annoyance, which may depend on and consequently affect their emotional state. In this paper we aim at investigating the interplay between the perceived sounds inside a vehicle and the driver's attention/enjoyment that can be improved through individualized acoustic-driven preference learning. The sound in the interior of a vehicle is influenced by several uncorrelated and dynamics factors such as road type, road quality, radio, and the number of passengers [8], to name a few. To tackle this problem, we can learn from the environment how to change the state inside the car to avoid an annoyance in the interior of the vehicle and keep driver's attention, closing the vehicle's windows or suggesting a different cruise speed.

Traditional learning techniques use an existing dataset and associated features to find trends in the data. Typically, this means the features need to be defined beforehand and once the input-output mapping is learned, and unless adapted to an online framework, remains stagnant. More recent advances in deep learning eliminate the need for identifying features beforehand but requires incredibly large amounts of data and computation time to learn the functional relationship between the data and output. Similar to the more traditional methods, these new artificial intelligence (AI) methods do not have an efficient updating scheme.

In this paper, we focus on developing/designing the learning framework for active vehicle adaptation based on psychoacoustic metric, derived from previous studies. The main advantage of our approach is in its capability of learning from a realistic simulation and not requiring annotated data. We evaluate the performance of the developed agent in a simulated environment without a user input. Future work will built upon these results to include user-specific adaptations.

II. Related Work

The safety of drivers has been improved over the decades through driver modeling studies [9]–[11] that broadened understanding of predicting driver's behaviors. Despite remarkable advances in visual perception, mostly applied in autonomous systems, and driver modeling, used in interactive systems, there is an important gap when considering the passengers, the driver and the vehicle as components of a vibro-acoustical system. This work builds upon findings from the areas of sound recognition and psychoacoustic sound evaluations. It has been recognized that driver's emotional state significantly affects the safety of driving.

Virtually all works on noise treatment in the vehicle have focused on noise detection or reduction for adjustments during design time of a vehicle. Several psychoacoustic indices have been introduced in sound quality evaluation engineering [12], [13] and evaluated in various driving scenarios. In the work of Nor et al. [14], for instance, both subjective and objective tests are studied to evaluate vehicle comfort index. They found that the metrics of loudness, sharpness, roughness and fluctuation strength are correlated with a human subject studies. Their experiments also showed how much the acoustical comfort is affected by each metric. Another interesting result of Nor et al.'s work is the study of the relation between road roughness (e.g., highway, smooth urban, dual carriage, pavement, and suburban) and the comfort index.

Duan et al. [15] studied vehicle interior noise under multiple working conditions: idle, constant speed, accelerating and braking. Duan et al. propose to predict the sound quality in the vehicle interior by using a neural network. They used four psychoacoustic parameters, the loudness, sharpness, articulation index, and A-weighted sound pressure levels (SPLs) as input features and subjective annoyance as a label for training the network. They created a dataset composed of 36 interior noise samples from the vehicle under the idle, constant speed, accelerating and braking conditions. The measurement method follows the GB/T 18697 standard. The results show an accuracy above 95.57%.

Although several psychoacoustic indices have been proposed in the past years, when driving, this information alone is not useful for identification of the sound source of annoyance. A straightforward approach is to separate the sound events and compute the annoyance of each one. Unfortunately, the detection and segmentation of general sound events is not still a well-solved problem.

The traditional pipeline composed of handcrafted features extraction such as Mel-frequency cepstrum (MFCC) and Spectrograms and general classifiers like SVM and GMM was extensively used to tackle sound recognition problems from speech recognition to scene classification and event detection. However, in the last few years, thanks to the advances in transfer learning techniques and cross-modal learning deep learning method attain the state-of-the-art results for most sound recognition problems. A challenging task in sound recognition is the classification of individual sounds in composite sounds. This task is called polyphonic sound event detection and consists of detecting multiple overlapping sound events. Several approaches have been proposed including transfer learning [16], cross-modal learning [17] and deep learning methods [18].

However, one particular problem of learning from sound samples is the lack of massive annotated data. This absence is particularly accentuated for sounds from the vehicle interior as most of the available sound datasets are composed of ambient, event or mixed sounds. Besides being influenced by several uncorrelated and dynamics factors, the sound in the interior of a vehicle also affects people (i.e., drivers) differently. Thus, in addition to the absence of data, a challenging task is how to learn to adapt from the environment and also from the drivers preferences to keep a pleasant environment inside the car for a particular driver.

Due to this scarcity of labeled data in sound classification tasks, several deep learning approaches build upon some transfer learning technique. Han et al. [16] tackle the absence of labeled data by combining active learning and a self-training approach to minimize the human annotation. Their methodology was able to provide a reduction of 52.2% in human labeled instances on the FindSounds dataset [1]. This dataset has nine categories and more than $16,930$ sound instances with durations ranging from 1 to 10 seconds.

SoundNet [17], for instance, provides a large-scale and rich representation of natural sound by using frames from labeled videos for learning a representation for raw audio waveforms. Despite using vision in training phase, the SoundNet only requires sound on test phase. Another multi-modal approach was presented by Ngiam et al. [19]. They demonstrated that better features could be learned from multiple modalities such as video and audio at learning time. The authors demonstrate the performance of the method on audio-visual speech classification task.

Hershey et al. [17] applied the most popular Deep Neural Networks (DNNs) used in image classification, i.e., AlexNet, VGG, Inception and ResNet in soundtrack classification task. They found that using embeddings features from these networks performs better than raw features in the AudioSet [13] (a dataset composed of over 1 million 10 second excerpts labeled acoustic events). The conclusion was that state-of-the-art image networks provide better results on audio classification than simple fully connected network or earlier image classification architectures such as AlexNet.

In a recent work, Çakir et al. [18] presented a convolution recurrent neural network (CRNN) for polyphonic sound event detection, where multiple overlapping sound events must be detected. After convoluting spectrograms with multiple convolutional layers and max-pooling, they fed a recurrent neural network (RNN), which has the output layers binarized for the event classification. Their experiments showed the proposed neural network outperformed traditional Gaussian Mixture Models-Hidden Markov Models (GMM-HMM) classifiers and methods using convolutional neural network or RNN.

Despite the impressive results on the supervised approaches, its full integration on real-world scenarios faces several challenges. Foremost, is the requirement of a large annotated dataset, which can be unfeasible to be created in some cases, where the set of labels is ill-defined or too large.

III. Methodology

In this section, we detail our exploration-exploitation approach based on a deep reinforcement learning, where an agent learns actions to perform in order to decrease the annoyance. Since the sound in the interior of a vehicle is influenced by several uncorrelated and dynamics factors, our agent can learn from the environment how to change the state inside the car to reduce the sound annoyance in the interior of the vehicle. Figure 1 depicts the main steps of the learning method.

[1]http://www.findsounds.com

A. Learning

In this work, we formulate the problem of choosing an action that avoids annoying sounds as a Markov Decision Process (MDP). Let $\mathcal{A} = \{a_1, \ldots, a_n\}$ be a discrete action set and $\mathcal{S}$ the state set. Thus, 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 a immediate reward r_i. In order to maximize the 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:

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

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

Our approach uses 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 select randomly an action and with probability $1-\epsilon$ we follow the action that maximizes we select the action that maximizes the quality of current and future actions. Thus, probability $1-\epsilon$ the action is determined by the policy

$$\pi(s_i) = \operatorname*{argmax}_a Q(s_i, a). \tag{3}$$

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 [20]:

- *Fluctuation* and *Roughness*: When we have in an environment multiple signals with different frequencies 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 a signal over the time. Fluctuation 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

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

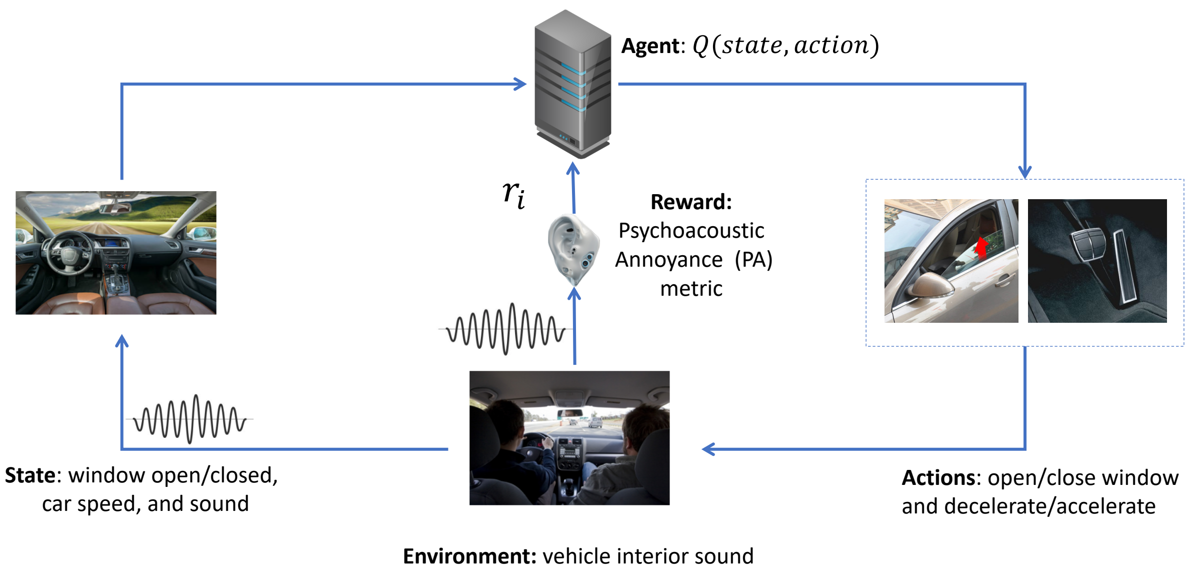

Fig. 1. Our approach is based on a deep reinforcement that learns from the environment how to change the state of the vehicle to avoid an unpleasant environment. After the agent took action a_i considering the environment state s_i, our system computes the reward r_i using the psychoacoustic annoyance (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.

A modulated signal is considerably more unpleasant when having a higher roughness and fluctuation;

- *Loudness*: The loudness is a psychological phenomenon based on the perceived loudness. Unlike the sound level that is a physical measurement, the loudness metric is based on subject studies using a group of people with normal hearing. In this study, in each person from the group listened to a tone at frequency f and a particular dB level. Then, another second tone with same dB level of the first one was played at a different frequency until it sounded equally as loud as the first tone. Formally, let E_{TQ} be the excitation at threshold in quiet, and E_0 be the excitation of the reference intensity, the specific loudness of a sound with excitation E is given by

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

The total loudness is the result of integrating the specific loudness over critical-band rates [2], i.e.,

$$N = \int_0^{24} N' dz, \quad (7)$$

where z is the critical-band in Bark.

- *Sharpness*: It is 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

$$S = \int_0^{24} S' dz, \quad (8)$$

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

The sound annoyance is closely related to the aforementioned psychoacoustic indices. Zwicker proposes to compute the psychoacoustic annoyance (PA) [20] as a function of sharpness, loudness, fluctuation, and roughness as follows:

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

where N_5 is the 95th percentile of loudness and

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

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

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

[2] A critical band is the smallest band of frequencies that activates the same part of the basilar membrane in the human hearing system.

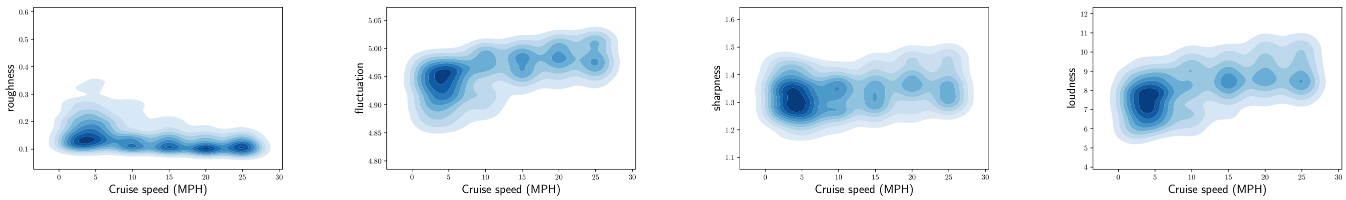

Fig. 2. Cruise speed versus psychoacoustic indices. Notice that while roughness decreases when increasing the cruise speed, fluctuation, sharpness, and loudness increase as cruise speed becomes greater. This relation creates a similar behavior on the psychoacoustic annoyance metric.

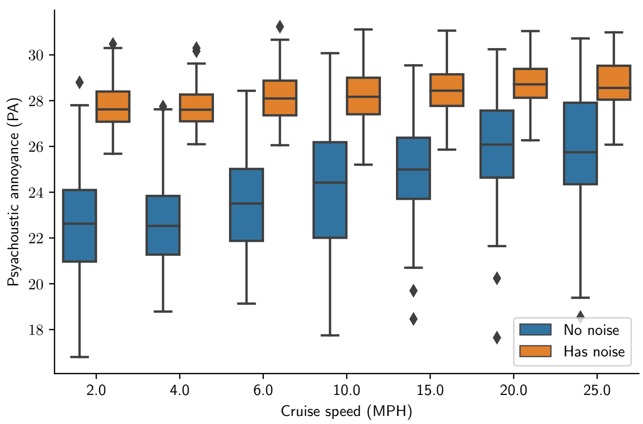

Fig. 3. The influence of the cruise speed on the PA metric. Increasing the cruise speed leads to higher PA value. The orange boxes are the values when including the sound of bells and beeps simulating noise outside from the vehicle. The diamonds represent observations outside $1.5\times$ interquartile range above the upper quartile and bellow the lower quartile.

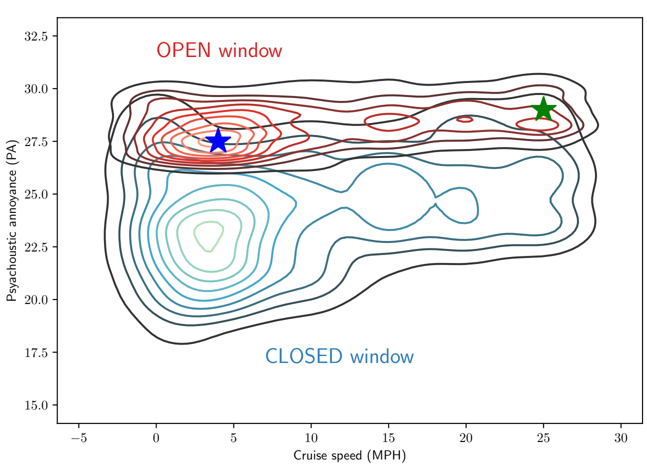

Fig. 4. Psychoacoustic annoyance metric measured in different configurations of the window status and cruise speed. The red curve depicts when the window is open and the blue curve when window is closed. All measures include additional noise (bells and beeps sound). Blue star represents a state where the agent can decrease the annoyance by simply closing the window. Green star is a more challenging state where the agent has to combine the actions of selecting a lower speed and closing the window to decrease the annoyance.

IV. Experiments

In this section, we present the experiments and the analysis of the effectiveness of our proposed framework. We used a simulator of a driving environment to create the soundscape, to explore and exploit the actions that minimize the annoyance.

A. Implementation details

We used a 5-layers neural network to represent the Q-function. The first layer is a fully connected layer of 128 neurons followed by a rectifier linear unit (ReLU) activation function. The second, third and fourth layers have 64, 32 and 24 neurons, respectively. A ReLU follows all of them. The last layer is composed of $|\mathcal{A}|$ neurons, producing the outputs as a weighted sum of the 4th layer followed by a sigmoid function.

The learning rate used was equal to 0.001. The exploration rate from 1 and to 0.1 with a decay rate of 0.95 and discount factor $\gamma = 0.8$. We set the memory size to 256 transitions, batch size 32 and 300 epochs. A window of size equal to 3 seconds is used to compute the psychoacoustic annoyance metric after performing the selected action.

B. Experimental Settings

We 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. Additionally to the sound environment of the simulator, we also included sounds of bells and beeps from outside in half of the time. The action set $\mathcal{A}$ is composed of 7 actions: changing the window state (open or closed) and six different cruise speed: 2, 4, 8, 10, 15 and 20 miles per hour.

In Figure 2 we can observe how the cruise speed affects the four psychoacoustic indices. Notice that as opposed to roughness that decreases when the cruise speed increases, all other indices, i.e., fluctuation, sharpness, and loudness increase as cruise speed becomes greater. The psychoacoustic annoyance metric has a similar behavior to fluctuation, sharpness, and loudness. As shown in Figure 3, the psychoacoustic annoyance is higher when the vehicle is running at higher speed.

Aside from using the sounds from the GTA simulator, we also included additional noise, i.e., sounds of bells and beeps. Figure 4 shows the values of psychoacoustic annoyance metric when setting different cruise speeds and changing the state of the vehicle's window. Whenever the window state is closed, the bells and beeps noise is not perceived by the agent. It can be seen that in some cases, as depicted by the blue star in Figure 4, it is sufficient the agent closes the window to

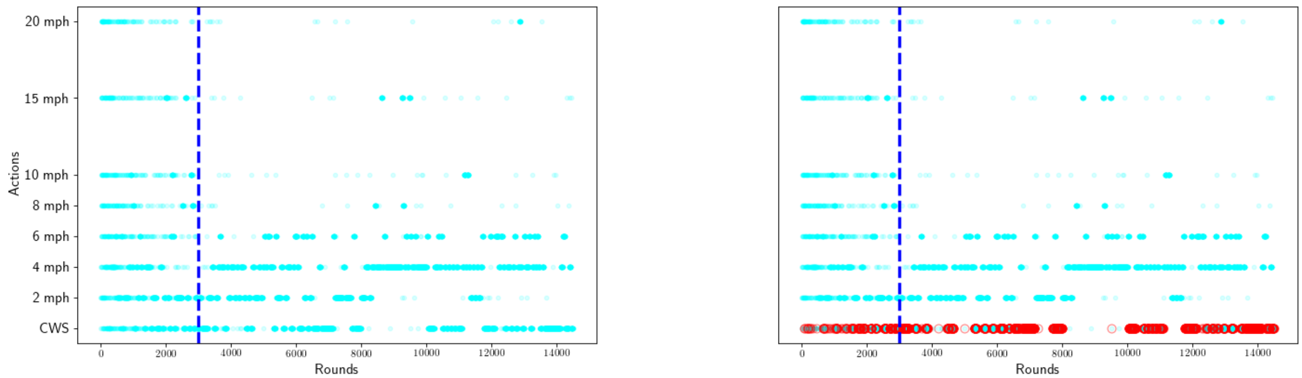

Fig. 5. On the left, we can see that in the beginning, the agent explores the action state by selecting different actions uniformly (CWS means Change Window State). After $3,000$ rounds (vertical blue line), the agent avoids choosing cruise speeds larger than 8 mph. On the right, the plot depicts the rounds where there was noise, and the window's state was open (red circles). In several rounds, the agent prefers closing the window when is perceived noise that is coming from outside of the vehicle.

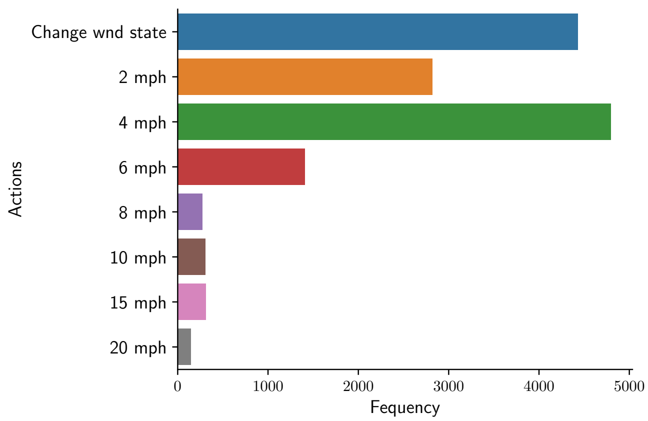

Fig. 6. The number of times the agent chose each action. These results show that, in general, agent learned to take actions that prevent large PA values.

decrease the annoyance. However, in other cases, like that one showed by the green star, the agent can only reduce the annoyance from the sound by combining actions of selecting a lower speed and closing the window.

C. Discussion

In Figure 5, the plot on the left shows the behavior of the agent selecting actions during the learning process. We can see that at the beginning of the training, the agent explores the action state by selecting different actions uniformly. After $3,000$ rounds (vertical blue line), the agent starts to avoid cruise speeds larger than 8 mph, since these speeds return large PA values.

More interesting, we can see that the agent not only learned to select low speeds but also learned to change the window state when it is necessary. On the right of Figure 5 we show red circles where there was noise, and the window's state was open. The first row on the plot represents the agent choosing to close the vehicle window to decrease the annoyance. It is worth noting that the agent correctly took the action that closes the window when there is noise from outside. Also, this behavior became predominant after $3,000$ rounds.

The bars in Figure 6 show the number of times the agent chose each action. As expected, most of the time the agent took actions to prevent large PA values, i.e., change window state and low cruise speed.

V. Conclusions

In this work, we present an acoustic-driven method based on deep reinforcement learning to adjust the interior vehicle state to improve driver's pleasantness. Our method measures the annoyance coming from the sound using psychoacoustic indices. This measurement is used to compute a reward that feeds a Q-network. The network encodes the best action to be taken to avoid annoyance as far as sound is concerned.

In our experiments, we showed the relation between the psychoacoustic indices and the cruise speed as well as the noise coming from outside of the vehicle. To simulate and controls this noise we add sound of bells and beeps. The results presented in this paper, all performed using the GTA V simulator, showed that the agent learned to choose actions to avoid creating annoying sounds.

As future works, we will further investigate other approaches, including agent as a recommendation system, where every time the driver follows a suggestion, the system is rewarded. Such reward can also be computed by measuring the behavior of the driver such as engaging in speeding. The second line of study will include an in-depth study of the influence of sounds on driver's habits, in particular, physiological signals. For the study, we will collect electroencephalogram (EEG), eye tracking data of the subjects while they perform the driving task.

REFERENCES

[1] J. Fagerlönn, "Urgent alarms in trucks: effects on annoyance and subsequent driving performance," *IET Intelligent Transport Systems*, vol. 5, no. 4, pp. 252–258, 2011.

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

[3] 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. 293–304, 2007.

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

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

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

[7] W. Krebber, M. Adams, F. Brandl, N. Chouard, K. Genuit, T. Hempel, R. v. Hofe, G. Irato *et al.*, "Objective evaluation of interior car sound œ the obelics project," 2000.

[8] P. Davies, "An overview of methods to quantify annoyance due to noise with application to tire-road noise," Citeseer, Tech. Rep., 2008.

[9] C. Little, "The intelligent vehicle initiative: advancing" human-centered" smart vehicles," *Public Roads*, vol. 61, no. 2, pp. 18–25, 1997.

[10] 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. 903–908.

[11] 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. 3461–3472, 2017.

[12] H. Fastl, "The psychoacoustics of sound-quality evaluation," *Acta Acustica united with Acustica*, vol. 83, no. 5, pp. 754–764, 1997.

[13] 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.

[14] M. J. M. Nor, M. H. Fouladi, H. Nahvi, and A. K. Ariffin, "Index for vehicle acoustical comfort inside a passenger car," *Applied Acoustics*, vol. 69, no. 4, pp. 343–353, 2008.

[15] 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.

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

[17] 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.

[18] 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. 1291–1303, 2017.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

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