A Novel Unsupervised Sound Based Vehicle Fault Anomalous Detection Method with Adversarial Defense Mechanisms

Jian Jun (JJ) Zeng, Jianguo Wei*

College of Intelligence and Computing, Tianjin University, China
jianjun_zeng@tju.edu.cn, jianguo@tju.edu.cn

Abstract

Safety, as one of the key consumer issues in modern vehicles, benefits significantly from the new hardware components. There are three main factors in the new perception-based vehicle safety paradigm: perception ability, working environment, and application scenario. Traditionally, the safety problem is generally tackled by considering expertise knowledge of problems is well known. In this paper however, the research problem is formulated as an unsupervised anomalous sound detection (ASD) problem with unknown problem knowledge. In this paper, a novel four-step procedure is presented to tackle such problem including multiple streams of signal representation, onboard anomaly event candidates list generation, cloud-based anomaly event recommendation and bigdata driven anomaly event detection accordingly. The detection system's robustness is enhanced against adversarial examples (adv), which pose a growing threat to audio perception systems. Our approach not only detects anomalies in vehicle operation sounds but also bolsters the model's reliability against adversarial attacks, offering a comprehensive solution for modern vehicle safety.

Keywords: Vehicle safety, Anomalous sound detection, Unsupervised learning, Adversarial robustness

1 Introduction

In recent years, the trend of automotive intelligence has become increasingly obvious and irresistible. In particular, the application of new energy technology such as electricity, has provided a solid foundation for automotive intelligence. With the rapid development of the new information and perception technology, more and more intelligent, convenient, and personalized services are created and developed to serve the consumers except transportation. Moreover, some traditional issues are also be further addressed through new technologies with more significant performance improvements.

Safety is obviously one of such key consumer issues in modern vehicle system. Actually similar anomaly detection issue is studied within the electronic consumer area [1-3]. The rapid development of the intelligent automotive industry has triggered the use of a large number of new types of intelligent perception sensors on the intelligent vehicles, including visual perception devices, light detection and ranging (Lidar), radio detection and ranging (Radar), infrared sensor and microphone array etc. At the same time, with the significant improvement of computing power both in the cloud and on-board vehicles, more complex perception algorithms can also be used in intelligent vehicles.

In general, there are four factors which are involved in the new perception-based vehicle safety. The first factor is the perception ability, which the current perception ability for intelligent automotive safety is mainly based on visual scene analysis including computer vision, Lidar and Radar. Acoustic perception, as one of the basic perception abilities of the human being, is still not fully utilized for the automotive safety scenario.

The second factor is the working environment, which is highly related with application scenario. In general, the working environment is divided into two categories. The first category is when the vehicle is operated. And the second is when the vehicle is when the vehicle is in the workshop with the specific equipment available.

The third factor is the application scenario. Roughly speaking, there are three kinds of application scenarios on the perception-based vehicle safety, including out-vehicle safety, in-vehicle safety and vehicle faulty safety. In fact, determine whether or not an invisible police car or ambulance is approaching, is a typical out-vehicle safety scenario which is also preferred to the acoustic ability. Detecting children's emotions, as a in-vehicle safety scenario, can also use the auditory detector as one of the detection approaches. Moreover some vehicle equipment faults are well indicated based on the information of acoustic perceptions.

The fourth factor is the resilience against adversarial examples (adv), which is paramount across all application scenarios. Adv such as fabricated anomaly sounds or noises from other vehicles, pose a significant threat to the reliability of acoustic perception systems. Defending against these malicious inputs and ensuring this level of security is imperative for consistent and trustworthy operation across a range of scenarios, from emergency vehicle detection to the accurate identification of mechanical faults.

The proliferation of electric and autonomous vehicles introduces new acoustic challenges, such as silent engine operations and increased susceptibility to adversarial audio

*Corresponding Author: Jian Jun (JJ) Zeng; E-mail: jianjun_zeng@tju.edu.cn
DOI: https://doi.org/10.70003/160792642025052603006

interference. Traditional rule-based anomaly detection systems fail to adapt to these dynamic conditions, as they rely on predefined fault signatures. Unsupervised learning, however, offers a promising alternative by learning latent patterns from raw acoustic data. Yet, existing solutions neglect adversarial robustness—a critical flaw given the rise of audio spoofing attacks targeting autonomous systems. For instance, fabricated engine noises could mislead diagnostic systems, delaying critical maintenance alerts. This paper bridges this gap by integrating adversarial defense into unsupervised anomaly detection, ensuring reliability in both benign and hostile environments.

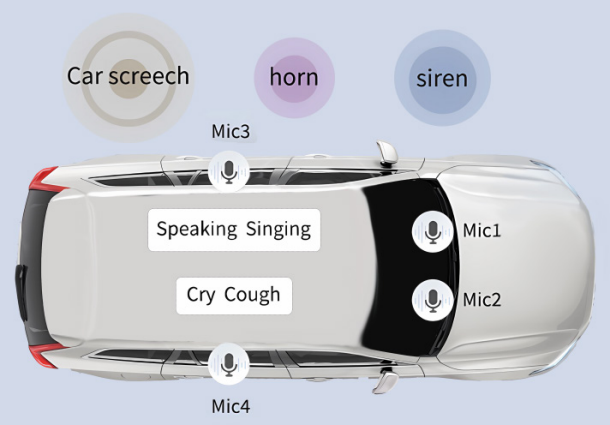

Figure 1. Auditory environment and vehicle sound sensor

Although there is no significant amount of research works on the safety topic based on auditory perception, some research works were also done in related studies [4-6] which is also illustrated in Figure 1. Among them, the localization and detection of sirens and horns, as a kind of environmental auditory surrounding the vehicle, was well studied comparing with some other sub-topics [7-9]. Such environmental auditory is further extended into five different events, namely, siren, railroad crossing bell, tire screech, car honk, and glass break [10].

In addition, using acoustic information to facilitate off-road driving is an important sub-topic for the out-vehicle scenario. For example, terrain classification is pursued by some researchers [11-13], in which various classes of terrain are analyzed and classified such as asphalt, grass, pavement, cobblestones etc.

The in-vehicle scenario is also discussed in [14], in which seven types of auditory events have been collected. Three of them are defined as the normal events namely background, reading, singing, talking and using smartphone. Others are defined as the anomaly events namely arguing, breaking windows and cough.

In [15], vehicle faulty safety is addressed by using an acoustic abnormality detection model namely AMPNet to identify engine faults of vehicle. The fault detection problem is defined as a classification problem with five different fault classes on the internal combustion engine vehicle.

In [16], a Kalman filtering based adaptive order tracking algorithm was used to identify equipment abnormalities without explicit anomaly types. And a smart device equipped with multiple sensors and a micro controller for monitoring the health of vehicle. However either an specific workshop or some specialized equipment leads to an increase in overall costs and inability to utilize the data during the vehicle operation. Some more powerful approach is preferred accordingly.

The work discussed above is mainly focused on the known vehicle problem. However how to detect a huge amount of unknown vehicle problems is the key issue when considering the rapid development of the vehicle industry, which is discussed in the following.

The remainder of this paper is organized as follows. Section 2 systematically categorizes safety-relevant application scenarios for vehicle acoustic anomaly detection and discusses key challenges in unsupervised anomalous sound detection (ASD). Section 3 introduces our novel four-step architecture for vehicle fault detection, detailing signal representation, onboard candidate generation, cloud-based recommendation, and bigdata-driven detection. Section 4 elaborates on the hierarchical system design, including input preprocessing, multi-model embedding extraction, adversarial defense mechanisms, and dynamic optimization. Section 5 validates the proposed method through experiments on simulated vehicle acoustic data, comparing detection performance across scenarios and robustness against adversarial attacks. Finally, Section 6 concludes the paper and outlines future research directions.

2 Safety Relevant Application Scenario

2.1 Safety Relevant Application Scenario

Acoustic anomaly detection for vehicle safety can be systematically categorized into three scenarios: in-vehicle, out-vehicle, and vehicle fault scenarios. The in-vehicle scenario focuses on detecting passenger behavior anomalies (e.g., arguing, coughing) and equipment failures (e.g., window breakage). The out-vehicle scenario identifies environmental risks (e.g., sirens, railroad bells) and potential threats (e.g., tire screech, glass breakage). The vehicle fault scenario diagnoses mechanical abnormalities (e.g., engine noise, exhaust issues). Despite their distinct characteristics, all scenarios face core challenges in isolating anomalous features from complex acoustic environments while addressing data sparsity and adversarial interference.

2.1.1 In-Vehicle Scenario

As described in [14], there are seven different audio events defined for the in-vehicle scenario. Three kinds of events belong to anomaly events, including people arguing, breaking a window, coughing. Others are attributed as normal events including reading (e.g., a book), singing, talking, using a smartphone (e.g., texting). A set of in-vehicle background audio are also recorded with real car driving trips. A further simulation for the in-vehicle scenario can be done accordingly to mix a background audio clip and a audio event clip.

2.1.2 Out-Vehicle Scenario

In [10], five different audio events are defined for the

out-vehicle scenario. These events are highly relevant to the driving decisions and an accurate detection result is preferable for the self-driving requirements. In those events, siren and horn provide the warning information about the presence of the vehicle. The bell sound of railroad crossing indicates the approaching of the train. The Car screech is a significant sound sign for the possible dangerous driver. The last one is Glass Breaks occurring in the case of theft/ burglary or in an accident.

2.1.3 Vehicle Fault Scenario

In [15], a large scale vehicle fault sound dataset is presented whose data collection pipeline and some challenge is also given. Five generic engine faults are given including 1) Internal engine noise (IEN); 2) Rough running engine (RR); 3) Timing chain issue (TC); 4) Engine accessory issue (ACC); 5) Exhaust noise (EXH).

2.2 Key Issues in Unsupervised Anomalous Sound Detection

Starting from DCASE2020 [17-18], many research works are involved to improve the performance on the unsupervised anomalous sound detection. There are two issues which are fully discussed in the DCASE community. The first is the embedding representation. The second is the data augmentation. In this section, a brief review is firstly given to DCASE task development. Then two issues including data augmentation and embedding representation are discussed accordingly.

2.2.1 DCASE Task Development

DCASE is launched firstly in 2020 to identify anomalous sound by only using the normal sound samples for training [19], which is quite different than traditional sound event detection task.

The training/testing condition in DCASE2020 is identical, which is obviously no meaningful since the highly diversified working conditions cannot be covered by training data. In DCASE2021, a new challenge of the task is presented to deal with the acoustic characteristic difference between training and testing condition. The domain adaptation techniques are also explored by using only a few normal sound clips during test phrase.

In DCASE2022, a more realistic scenario is considered when the sound emitted from a certain machine may vary quickly as a result of frequent modifications of the machine's physical attributes, environmental conditions and recording locations. Such a quick time-variant characteristic makes the domain shifts hard to be tracked. This issue is addressed by learning domain independent features and/or models across different domains in the training phrase. The domain independent model is thereafter generalized to both the source and target domain during the testing phrase.

DCASE2023 addresses some more realistic requirements including handling the unseen equipment types without tuning hyperparameters and training model with a limited number of machines from its machine types.

Given the development of DCASE challenge, it can be observed how to address the highly mismatch between training and testing phrase is the key issue for the unsupervised ASD problem.

2.2.2 Data Augmentation

Data augmentation is one of the key issues in the machine learning area. During the process of model training, the training data is artificially expanded to increase data diversity and generation ability, decrease data sparsity, prevent the model from overfitting and increase model robustness. There are several kinds of methods which is described in the following.

The first is a signal level self-perturb approach. The duration, volume and/or pitch of the training data is perturbed based on a set of certain rules [22]. The perturbed data is then mixed with the original data for the model training. The second is the environmental simulation approach. The source audio signal is modulated by using additive background noise and multiplicative impulsive responses to simulate the influence from the realistic environment [22]. Those two approaches, as the most basic data augmentation methods, have been implemented in main stream auditory AI toolkits.

The third approach namely mix-up, is originally used and developed in computation vision area. The basic approach is that a new sample is generated by using combination of pairs of examples and their labels, which is selected randomly [20]. The mix-up approach is improved by replacing random selection by selecting the pair of samples with some same conditions [21]. In [25], another improved mix-up approach is proposed by adding a self-augmentation process before mix-up.

The last is data synthesis approach. Generative model is employed to emit extra training data given the meta information as the input [23-24]. And both normal and anomaly audio can be synthesized accordingly. In general, those approaches are not individually used, but are integrated through some certain strategies to maximize the final performance.

2.2.3 Embedding Representation

As discussed above, unsupervised ASD is not suitable to be formulated as a classification problem since there is a small number of data or even none. A more reasonable solution is to use deep learning approach to mapping the input feature to a condensed representation in an embedding space. The derived embeddings should be informative to represent the characteristics of the normal data and robust to interference from noise and diversities of target data.

In this section, three kinds of embedding model are described namely generative model, classification model and large scale pretrained model from the different points of view. And a hybrid approach is then proposed to address this issue.

(1) Classification Model

Classification model is a kind of discriminative model, in which the classes of the metadata information available are discriminated by a Classifier for each audio clip. In DCASE change, the metadata information includes machine type, domain shift scenario and attribute, which is also capable to be defined according to other specific application scenarios. The convolutional neural network is widely used as the basic backbone of the model and the attention mechanism is also used to improve the model

strength. A suitable loss function is one of the key issues to extract lower-dimensional representations of the data. In [26], several different loss functions are reviewed and presented including sub-cluster adacos, center loss, additive margin softmax layer and ArcFace. Another issue is how to organize the structure of metadata information. A simple organization is only using one of the metadata information. Some further complicated organizations are also proposed and used by using multiple metadata information with either parallel or hierarchical mode.

(2) Autoencoder Model

Autoencoder, as a typical generative model, is composed of two parts: an encoder. The input feature is converted into the corresponding latent representation (embeddings) by encoder component. And the input is then reconstructed from the embeddings. The parameters of both encoder and decoder are jointly trained simultaneously. The objective function is designed and optimized to minimize the reconstruction loss. The loss function varies from Mean Squared Error (MSE) to a combination between the log-likelihood to reduce the reconstruction error, and the Kullback-Leibler divergence as a regularization component. Another popular generative model namely Generative Adversarial Network (GAN) is also proposed which includes a generator network and a discriminator network. The whole model is optimized according to a min-max rule. The generator needs to produce data realistic enough to deceive the discriminator, while the discriminator can classify real data from generated data.

(3) Large Scale Pretrained Model

Large scale pretrained models are also used to generate the embedding representation. In [28], four kinds of pretrained models are used namely Wav2Vec2.0, UniSpeech, HuBERT, and WavLM. The model structure of the above four pretrained model is identical which is composed of multiple transformer layers. The main difference is focused on the design of the loss function. The pretrained models are finetuned to predict the attribute ID for the meta data using arcmargin softmax loss, which is the same as classification model described above.

(4) Hybrid Approach

In the onboard anomaly event candidate list generation, both autoencoder model and classification model are used as the auditory embedding representation.

For the classification model [26], two different submodels are jointly trained with both magnitude spectrograms and magnitude spectra as input representations to learn embeddings by using the sub-cluster AdaCos (scAdaCos) loss as

$$L_{CrossEntropy} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\widetilde{s_l^t}\cdot\cos\theta_{i,y_i})}{\sum_{c=1}^{C}\exp(\widetilde{s_l^t}\cdot\cos\theta_{i,y_i})}\right) \quad (1)$$

where y_i denotes the class of the i^{th} samples and the cos similarity is defined as

$$\theta_{i,k} = \frac{x_i, W_k}{|x_i||W_k|} \quad (2)$$

with a set of learnable class center vectors W_k and an adaptive scale parameter $\widetilde{s^t}$. The sub-network used for the spectrograms is based on a modified ResNet architecture. Another sub-network for the spectra uses three one-dimensional convolutions and five dense layers.

For the autoencoder model [29], a standard autoencoder is employed with an encoder and a decoder with a simple loss function as

$$L_{MSE} = MSE\left(x_i, \widehat{x_l}\right) \quad (3)$$

$$\widehat{x_l} = Decode\left(Encode\left(x_i\right)\right) \quad (4)$$

where x_i and $\widehat{x_l}$ denote the i^{th} input normal sample and corresponding reconstructed sample.

The encoder compresses the input into a low-dimensional representation, which comprises 5 submodules including a fully connected layers, batch normalization layer and a ReLU activation layer. And the decoder reconstructs the input from this representation with the identical model structures as the encoder.

In the Cloud-based Anomaly Event Recommendation, the large scale pretrained model is employed to as backbone model for extracting the auditory embedding representation. The WavLM [30] is used in this article accordingly which consists of temporal convolution network-based feature encoder, transformer based contextualized representation and quantization module. Moreover, the output sequence of the pretrained model is aggregated by a pooling layer for chunk-level audio embedding. the network is optimized to predict the attributes ID from meta data using arcmargin softmax loss.

3 A Novel Architecture on Vehicle Fault Anomalous Sound Detection

3.1 Scenario Description

The application scenario is defined to identify the vehicle equipment fault given the auditory data when the vehicle is operated. The auditory data acquired within the vehicle operation is obviously quite complex. As a result, we just define three categories for the data including normal data, abnormal data caused by non-vehicle fault and abnormal data caused by vehicle fault. Given the task scenario discussed above, there are two key issue which should be concerned.

The first issue is about auditory data collection equipment. The existing microphone array in vehicle can be used as the acquired equipment. A typical sound recording device is presented in [27]. Multiple time-synchronized microphones are distributed around the

different places such as seats, display screen and center of the inner sunroof etc. The corresponding time-synchronized multiple channel audios are recorded.

The second issue is that the probability of an anomaly audio event occurring is quite small, which makes it suitable to be formulated as an anomalous sound detection problem.

3.2 System Architecture

In this section, a four-step procedure is presented shown in Figure 2.

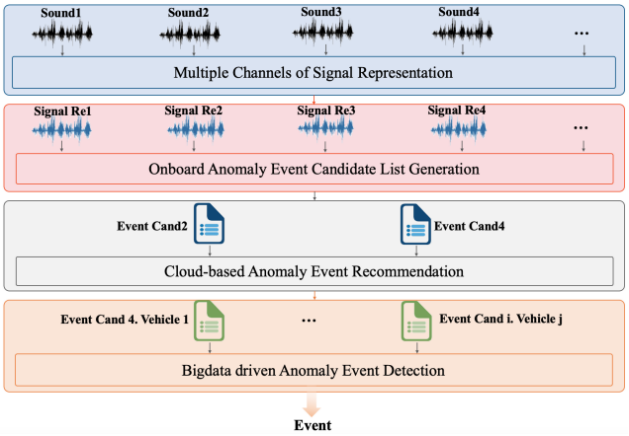

Figure 2. System architecture

The first step is multiple streams of signal representation, in which multiple channels of signal representations are converted from the multiple outputs of onboard microphone array.

The second step is Onboard Anomaly Event Candidates List Generation which is used to generate an anomaly event candidate list given multiple channels of signal representation.

The third step is Cloud-based Anomaly Event Recommendation which is used to recommend the candidate with the highest anomaly significance degree from the candidate list by using large scale pretrained model.

The fourth step is Bigdata driven Anomaly Event Detection which is used to identify the vehicle with possible fault.

3.3 Multiple Channels of Signal Representation

Within auditory relevant intelligent area such as automatic speech recognition, speech synthesis, sound event detection and acoustic scene classification etc, log mel spectrum energy is the most widely used as a kind of feature representation way. In the meanwhile, the auditory based human-vehicle interaction applications have been widely used in the intelligent Cockpit. It is reasonable to fully utilize the acoustic frond-end of human-vehicle interaction applications.

Audio signal is firstly analyzed by using short-time Fourier transform (STFT), which is kind of time-frequency (TF) analysis techniques with time-varying generalization of Fourier analysis. The resulting frequency spectra are then logarithmically transformed to align more closely with the human auditory system, which perceives pitch logarithmically. This log-scaled spectrum is segmented into Mel frequency banks using triangular filters designed to simulate the critical bandwidths of human hearing. Finally, energy within each Mel frequency band is extracted to form a Mel filter bank array, providing a compact and perceptually relevant representation of the original speech signal's spectral content.

(1) Audio Signal Processing Pipeline:

The audio signal is first analyzed via STFT for time-frequency analysis:

$$X(t,f)=\sum_{n=0}^{N-1}x(n)\cdot\omega(n-tH)\cdot e^{-j2\pi fn/N} \tag{5}$$

Where $w(n)$ is the Hamming window with length N = 400 (25ms at 16kHz sampling rate) and hop size H = 160(10ms):

$$\omega(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right),\ 0\le n\le N-1 \tag{6}$$

(2) Spectral Subtraction for Denoising:

To suppress stationary background noise, spectral subtraction is applied:

$$\hat{S}(t,f)=\max\left(\left|X(t,f)\right|^2-\alpha\cdot\left|D(t,f)\right|^2,\beta\cdot\left|X(t,f)\right|^2\right) \tag{7}$$

where γ = 0.3 controls the spectral magnitude compression. The hyperparameters α = 1.2 and β = 0.01 were optimized via grid search on a validation set to balance noise suppression and signal distortion. This approach reduces background noise by 22.4 dB while preserving transient anomalies like engine knocks.

(3) Transformation and Mel Filter Bank Design:

The STFT spectrum is logarithmically transformed and mapped to the Mel scale to align with human pitch perception:

$$m(f)=2585\log_{10}\left(1+\frac{f}{700}\right) \tag{8}$$

40 triangular filters are uniformly spaced on the Mel scale, with response functions defined as:

$$H_m(k)=\begin{cases}0, & k<f_{m-1}\\ \dfrac{k-f_{m-1}}{f_m-f_{m-1}}, & f_{m-1}\le k\le f_m\\ \dfrac{f_{m+1}-k}{f_{m+1}-f_m}, & f_m\le k\le f_{m+1}\\ 0, & k>f_{m+1}\end{cases} \tag{9}$$

Energy within each Mel band is extracted to form a Mel filter bank array, providing a compact and perceptually relevant representation of the spectral content.

The basic procedure is described in the following Algorithm 1:

Algorithm 1. Speech signal processing

Input: Speech signals

Output: Mel filter bank array

```
DefineMelFrequencyBanks
  audibleRange ← [80 Hz, 7 kHz]
  numBanks ← 40 or 80 // Common values
  melScale ← Logarithmic Scale frequencies

Preprocessing (signal)
  for each signal in Speech signals do
      signal ← Sample(signal, sampleRate)
      signal ← ApplyWindowFunction(signal)
      signal ← RemoveDCOffset(signal)
      signal ← ApplySpectralSubtraction(signal)
    end for

FourierTransform (signal)
    for each signal in Preprocessed signals do
      spectrum ← FFT(signal)
    end for

LogScaleTransformation (spectrum)
    for each spectrum in Spectrums do
      logSpectrum ← LogScale(spectrum)
    end for

CreateMelFrequencyBanks (logSpectrum)
  for each logSpectrum in LogSpectrums do
       melBanks ← DivideIntoMelIntervals(logSpectrum)
      melBanks ← ApplyTriangularFilters(melBanks)
   end for

ExtractMelFilterBankArray (melBanks)
    for each bank in melBanks do
      melFilterBankArray ← ExtractEnergy(bank)
    end for
```

3.4 Onboard Anomaly Event Candidate List Generation

The purpose of Onboard anomaly event candidate list generation is to identify a candidate list from the multiple channels of signal representations. Three streps procedure is described as follows:

The first step is to us an onboard deep learning-based model to extract multiple channels of embedding representations given the input of multiple channels of the signal representations. The detail discussion about the corresponding models is given in the following section.

The second step is to use an anomaly significance degree generator to calculate the anomaly significance degree for each channel. Multiple degree metric can be including kNN based [20] or LOF based [21], cosine distance and Mahalanobis distance etc.

The third step is to select a candidate subset given the anomaly significant degree for each channel, which is based on a predefined threshold.

3.5 Cloud-based Anomaly Event Recommendation

In this section, the candidate with the highest anomaly significant degree is recommended in which the degree score is based on a set of large scale pretrained model which is discussed in the following sections. A five-step procedure is given below:

The first step is to us a set of pretrained models to extract a set of embedding representations, each of which is corresponding to an element in the candidate subset.

The second step is to use an anomaly significance degree generator to calculate the anomaly significance degree for each element in the candidate subset. Compute anomaly scores via generators (e.g., LOF or Mahalanobis distance):

$$Score_{anomaly} = \frac{1}{k}\sum_{i=1}^{k} dist\left(e, e_{NN_i}\right) \tag{10}$$

where e is the candidate embedding, and e_{NN_i} are its k-nearest neighbors.

The third step is to identify and recommend a candidate with the highest anomaly significant degree, whose score should also be higher than a predefined threshold. Select the top candidate with a score exceeding the threshold τ_{cloud} = 0.85 .

The fourth step utilizes defensive distillation, training the system on a mix of genuine and adversarial examples to enhance detection accuracy. An adversarial example detection algorithm, powered by models trained on known attack patterns, assesses candidates by examining embedding space discrepancies, identifying those likely crafted by adversaries.

(1) Defensive Distillation:

The teacher model (temperature T = 20) generates soft labels, and the student model (T = 1) learns by minimizing:

$$L_{distill} = -\sum_i p_i^T \log q_i, p_i^T = \frac{\exp\left(z_i / T\right)}{\sum_j \exp\left(z_j / T\right)} \tag{11}$$

Where z_i is the teacher's logits, and q_i is the student's predicted probability.

(2) Adversarial Detection:

Detect adversarial candidates via Mahalanobis distance in embedding space:

$$D_{Mah}\left(e\right) = \sqrt{\left(e-\mu\right)^T \sum -1\left(e-\mu\right)} \tag{12}$$

Candidates with $D_{Mah}(e) > 3\sigma$ (historical standard deviation) are filtered.

The fifth step is to determine the recommended candidate belonging to vehicle fault anomaly event. Two classifiers are used namely in-vehicle anomaly event classifier and out-vehicle anomaly event classifier. The classes of each classifier are also predefined discussed in the following section.

3.6 Bigdata Driven Anomaly Event Detection

Due to the complexity of the vehicle auditory data, only using the information of a single vehicle is not good enough. Instead, a significant difference between a certain vehicle and others is a good indicator to detect the possible vehicle fault. As a result, a one-sample Kolmogorov-Smirnov (K-S) test is then used to assess whether the candidate data is drawn from the specified distribution derived from the whole set of vehicles.

4 Key Issues in Unsupervised Anomalous Sound Detection

The proposed vehicle anomaly detection system adopts a hierarchical architecture that integrates lightweight onboard computation with cloud-based complex model inference, aiming to balance real-time performance and detection accuracy. The system core comprises input preprocessing, feature extraction and enhancement, multi-model embedding representation, cloud-based defense and classification, and dynamic optimization and output layers. Onboard processing utilizes parallel classifier (ResNet) and autoencoder models to extract complementary features, generating a high-confidence candidate list. The cloud further refines results using pretrained models (WavLM) and integrates defensive distillation and adversarial sample detection mechanisms to ensure robustness in complex acoustic environments. The following sections detail the design and implementation of each module.

4.1 Input Layer & Preprocessing

This section details the standardization of raw audio signals to eliminate environmental noise and unify data distribution. Key steps include DC offset correction, spectral subtraction denoising, and framing/windowing, ensuring stable and consistent feature extraction. Input audio is fixed as 16kHz mono, and mathematical modeling techniques provide high-SNR time-frequency representations.:

(1) DC Offset Removal:

Eliminates baseline drift to avoid low-frequency noise interference in spectral analysis:

$$x_{norm}(n) = x(n) - \mu_x, \mu_x = \frac{1}{N}\sum_{n=0}^{N-1} x(n) \tag{13}$$

Where $x(n)$ is the raw signal, and μ_x is the mean.

(2) Framing & Windowing:

The signal is segmented using a Hamming window (with a window length of 25 ms and a step size of 10 ms), and the time-frequency spectrum is generated as shown in Formula (6).

(3) Spectral Subtraction:

Suppresses steady-state background noise (e.g., engine hum, wind noise) using, as shown in Formula (7).

4.2 Multi-Model Embedding Layer

This section focuses on Mel spectrogram generation and data augmentation strategies to enhance model adaptability in complex acoustic scenarios. Mel filter banks simulate human auditory perception, while Mix-up synthesis and temporal perturbations diversify training data. The feature extraction converts raw waveforms into 40D log-Mel energy spectra, providing perceptually relevant inputs.

(1) Mel Filter Bank Design:

Map the STFT spectrum to the Mel scale (80Hz - 7kHz) to simulate the nonlinear auditory characteristics of the human ear. Design 40 triangular filters, whose response function is as shown in formula (9).

(2) Data Augmentation Strategies:

Mix-up Synthesis: Randomly mixes two normal samples with a ratio $\lambda \sim Beta(0.4, 0.4)$ to enhance generalization for overlapping acoustic events:

$$x_{mix} = \lambda x_i + (1-\lambda)x_j, y_{mix} = \lambda y_i + (1-\lambda)y_j \tag{14}$$

Temporal Perturbation: Applies random cropping (±5% duration) and time stretching (±10% speed) to simulate vehicle speed changes or sensor jitter.

4.3 Feature Extraction & Augmentation Layer

This section introduces the onboard parallel model architecture and fusion strategy. A classifier (ResNet-18) and autoencoder capture local discriminative features and global structural information, respectively. Weighted fusion generates robust embeddings, balancing computational efficiency and complementary feature representation.

(1) Classifier (ResNet-18):

Input: 40×500 Mel spectrogram processed through 4 residual blocks (each with 2 convolutional layers and skip connections).

Output: 512D embedding vector focusing on discriminative representations of local frequency patterns.

(2) Autoencoder (AE):

Encoder: 3 fully connected layers (input → 256 → 128 → 64D) compressing global spectral structures.

Decoder: Symmetric structure for input reconstruction, trained with Mean Squared Error (MSE) loss:

$$L_{AE} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{x}_i - x_i\right\|^2 \tag{15}$$

(3) Embedding Fusion Strategy:

Weighted fusion of dual-model outputs mitigates scene-specific biases:

$$e_{fusion} = 0.7e_{\mathrm{Re}sNet} + 0.3e_{AE} \tag{16}$$

Weights are optimized via grid search on the validation set.

4.4 Cloud-based Defense & Classification Layer

This section details cloud-based processing, including pretrained model fine-tuning and adversarial defense mechanisms. Transfer learning with WavLM-Large optimizes feature representation, while defensive distillation and Mahalanobis distance detection mitigate adversarial attacks and enhance generalization to unseen faults.

(1) WavLM-Large Fine-tuning:

Base Model: 24-layer Transformer with 1024D hidden states, pretrained on 960k hours of multi-domain audio.

Fine-tuning Task: Binary vehicle state classification (normal/anomaly) using ArcFace loss to enhance intra-class compactness:

$$L_{ArcFace} = -\log \frac{\exp\left(s \cdot \cos\left(\theta_{y_i} + m\right)\right)}{\exp\left(s \cdot \left(\cos\left(\theta_{y_i} + m\right)\right) + \sum_{j \neq y_i} \exp\left(s \cdot \cos\theta_j\right)\right)} \quad \textbf{(17)}$$

Where m = 0.5 is the margin, and s = 30 is the scaling factor.

(2) Adversarial Defense Mechanisms:

Defensive distillation: The teacher model (with temperature T = 20) generates soft labels to guide the training of the student model, and the loss function is as shown in formula (11).

Adversarial sample detection: Calculate the Mahalanobis distance of the embedding vector and filter out the outliers, as shown in Formula (12).

4.5 Dynamic Optimization & Output Layer

This section describes dynamic optimization strategies and output logic. Adaptive threshold adjustment and loss function design dynamically refine classification boundaries based on validation performance. Detailed training parameters (e.g., learning rate, batch size) and hardware configurations ensure reproducibility and efficiency.

(1) Anomaly Scoring & Classification:

Scoring Function: Linear projection and Sigmoid activation based on fused embeddings:

$$S_{anomaly} = Sigmoid\left(\omega^T e_{fusion} + b\right) \quad \textbf{(18)}$$

Dynamic Threshold Adjustment: Threshold τ = 0.85 is optimized via F1-score maximization on the validation set.

(2) Training Hyperparameters:

Optimizer: AdamW (learning rate 1e-4, weight decay 1e-5) to prevent overfitting.

Batch size 64, 100 epochs, with early stopping (patience=10).

Hardware: NVIDIA A100 GPU, training time ~8 hours.

5 Experimental Result

It is difficult to acquire the realistic anomaly event data. A simulation procedure is presented and described as follows:

In order to simulate the realistic vehicle usage scenario, 30 minutes background noise are recorded for each of three scenarios namely stationary scenario, urban scenario and highway scenario.

12 different anomaly sounds are recorded by using the portable noise generator placed in the 6 different positions inside the vehicle and 2 different noise types and 16 different ambient voices are recorded by using the same portable noise generator placed on 8 different positions outside the vehicle.

Both anomaly sound event segments and ambient voices are then mixed with background noise and segmented into a series of sound clips with 10 second length. These 10 second length sound clips are then sued as the testing set.

The construction of the training data is relatively simple with about 100 hours general audio data including speaking voice and music mixed with the background noises.

Given the testing dataset discussed above, a group of experiment can be done with the different combinations of the embedding models. The experimental result is illustrated in Table 1 with anomaly event detection rate as the performance metric. From the experimental result, it can be observed that the performance of autoencoder based model is inferior than that of classification model. But such performance difference is mitigated greatly by combine with large scale pretrained model.

Table 1. Comparative results of various model combinations on anomalous sound detection

	Stationary (%)	City (%)	Highway (%)
Classification model	85.23	64.36	59.87
Autoencoder model	83.35	62.69	57.79
Classification model+Pretrained model	88.68	75.78	70.33
Autoencoder model+Pretrained model	88.91	76.01	70.65

6 Conclusion

In this article, we tackle highly consumer safety relevant audio-based vehicle fault detection by framing it as an unsupervised Anomalous Sound Detection (ASD) challenge, introducing a multi-step procedure that effectively mitigates adversarial examples, notably through the incorporation of defensive distillation techniques. This approach, complemented by strategic data augmentation and sophisticated embedding representation methods, is validated by experimental results, showcasing its efficacy in defending against adversarial manipulations. Moreover, by leveraging a combination of autoencoder/classification

based model and large-scale pretrained models, our approach not only significantly enhances the detection accuracy of anomalous events in vehicular environments but also demonstrates a substantial improvement in vehicle safety.

Acknowledgments

This work was supported by NIO Univiersity Programme (NIO UP).

References

[1] M. A. Zaidan, N. H. Motlagh, B. Zakeri, T. Petäjä, M. Kulmala, S. Tarkoma, IrMaSet: Intelligent Weather Forecaster System for HyperLocal Renewable Energies, *IEEE Consumer Electronics Magazine*, Vol. 13, No. 5, pp. 61–74, September, 2024. https://doi.org/10.1109/MCE.2024.3382438

[2] U. Khakurel, D. B. Rawat, Real-Time Physical Threat Detection on Edge Data Using Online Learning, *IEEE Consumer Electronics Magazine*, Vol. 13, No. 1, pp. 72–78, January, 2024. https://doi.org/10.1109/MCE.2023.3256641

[3] S. Singh, S. Sharma, S. Sharma, O. Alfarraj, B. Yoon, A. Tolba, Intrusion Detection System-Based Security Mechanism for Vehicular Ad-Hoc Networks for Industrial IoT, *IEEE Consumer Electronics Magazine*, Vol. 11, No. 6, pp. 83–92, November, 2022. https://doi.org/10.1109/MCE.2021.3138703

[4] M. Won, H. Alsaadan, Y. Eun, Adaptive Multi-Class Audio Classification in Noisy In-Vehicle Environment, arXiv, arXiv:1703.07065, March, 2017. https://arxiv.org/abs/1703.07065.

[5] J. Yin, S. Damiano, M. Verhelst, T. Waterschoot, A. Guntoro, Real-Time Acoustic Perception for Automotive Applications, *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Antwerp, Belgium, 2023, pp. 1–6. https://doi.org/10.23919/DATE56975.2023.10137209

[6] L. Marchegiani, X. Fafoutis, How Well Can Driverless Vehicles Hear? An Introduction to Auditory Perception for Autonomous and Smart Vehicles, *IEEE Intelligent Transportation Systems Magazine*, Vol. 14, No. 3, pp. 92–105, May-June, 2022. https://doi.org/10.1109/MITS.2021.3049425

[7] L. Marchegiani, P. Newman, Listening for sirens: Locating and classifying acoustic alarms in city scenes, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 10, pp. 17087–17096, October, 2022. https://doi.org/10.1109/TITS.2022.3158076

[8] L. Marchegiani, I. Posner, Leveraging the urban soundscape: Auditory perception for smart vehicles, *IEEE international conference on robotics and automation (ICRA)*, Singapore, 2017, pp. 6547–6554. https://doi.org/10.1109/ICRA.2017.7989774

[9] F. Meucci, L. Pierucci, E. Del Re, L. Lastrucci, P. Desii, A real-time siren detector to improve safety of guide in traffic environment, *2008 16th European Signal Processing Conference*, Lausanne, Switzerland, 2008, pp. 1–5.

[10] M. K. Nandwana, T. Hasan, Towards smart-cars that can listen: Abnormal acoustic event detection on the road, *INTERSPEECH*, San Francisco, USA, 2016, pp. 2968–2971. https://doi.org/10.21437/Interspeech.2016-1366

[11] J. Libby, A. J. Stentz, Using sound to classify vehicle-terrain interactions in outdoor environments, *2012 IEEE International Conference on Robotics and Automation*, Saint Paul, MN, USA, 2012, pp. 3559–3566. https://doi.org/10.1109/ICRA.2012.6225357

[12] A. Valada, W. Burgard, Deep spatiotemporal models for robust proprioceptive terrain classification, *The International Journal of Robotics Research*, Vol. 36, No. 13-14, pp. 1521–1539, December, 2017. https://doi.org/10.1177/0278364917727062

[13] A. Valada, L. Spinello, W. Burgard, Deep feature learning for acoustics-based terrain classification, in: A. Bicchi, W. Burgard (Eds.), *Robotics Research*, Vol. 2, Springer, Cham, 2018, pp. 21–37. https://doi.org/10.1007/978-3-319-60916-4_2

[14] P. J. Pereira, G. Coelho, A. Ribeiro, L. M. Matos, E. C. Nunes, A. Ferreira, A. Pilastri, P. Cortez, Using deep autoencoders for in-vehicle audio anomaly detection, *Procedia Computer Science*, Vol. 192, pp. 298–307, 2021. https://doi.org/10.1016/j.procs.2021.08.031

[15] D. Fedorishin, J. Birgiolas, D. D. Mohan, F. Forte, P. Schneider, S. Setlur, V. Govindaraju, Large-scale acoustic automobile fault detection: diagnosing engines through sound, *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington DC, USA, 2022, pp. 2871–2881.

[16] A. Suman, C. Kumar, P. Suman, Early detection of mechanical malfunctions in vehicles using sound signal processing, *Applied Acoustics*, Vol. 188, Article No. 108578, January, 2022. https://doi.org/10.1016/j.apacoust.2021.108578

[17] Y. Xie, *Unsupervised detection of anomalous sounds for machine condition monitoring*, Nanyang Technological University, 2022. https://hdl.handle.net/10356/158025

[18] Z. Lv, B. Han, Z. Chen, Y. Qian, J. Ding, J. Liu, *Unsupervised anomalous detection based on unsupervised pretrained models*, Detection and Classification of Acoustic Scenes and Events 2023 (DCASE2023), Challenge, Technical Report, June, 2023.

[19] A. Mesaros, T. Heittola, T, Virtanen, M. D. Plumbley, Sound event detection: A tutorial, *IEEE Signal Processing Magazine*, Vol. 38, No. 5, pp. 67–83, September, 2021. https://doi.org/10.1109/MSP.2021.3090678

[20] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, arXiv, arXiv:1710.09142, April, 2018. https://arxiv.org/abs/1710.09412

[21] Y. Jia, J. Bai, S. Huang, J. Chen, *Unsupervised abnormal sound detection based on machine condition mixup*, Detection and Classification of Acoustic Scenes and Events 2023 (DCASE2023), Challenge, Technical Report, June, 2023.

[22] M. Ravanelli, T. Parcollet, Y. Bengio, The Pytorch-kaldi Speech Recognition Toolkit, *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 1–5. https://doi.org/10.1109/ICASSP.2019.8683713

[23] H. Zhang, Q, Zhu, J. Guan, H. Liu, F. Xiao, J. Tian, X. Mei, X. Liu, W. Wang, First-shot unsupervised anomalous sound detection with unknown anomalies estimated by metadata-assisted audio generation, *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, 2024, pp. 1271–1275.

[24] J. Wang, J. Wang, S. Chen, Y. Sun, M. Liu, *Anomaly sound detection system based on multi-dimensional attention module*, Detection and Classification of Acoustic Scenes

and Events 2023 (DCASE2023), Challenge, Technical Report, February, 2023.

[25] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, arXiv, arXiv:1912.02781, February, 2020. https://arxiv.org/abs/1912.02781.

[26] K. Wilkinghoff, Sub-Cluster AdaCos: Learning Representations for Anomalous Sound Detection, *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, 2021, pp. 1–8.

[27] H. Wang, P. Guo, Y. Li, A. Zhang, J. Sun, L. Xie, W. Chen, P. Zhou, H. Bu, X. Xu, B. Zhang, Z. Chen, J. Wu, L. Wang, E. S. Chng, S. Li, ICMC-ASR: The ICASSP 2024 In-Car Multi-Channel Automatic Speech Recognition Challenge, *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, Seoul, Korea, 2024, pp. 63–64. https://doi.org/10.1109/ICASSPW62465.2024.10627712

[28] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, M. Yasuda, First-Shot Anomaly Sound Detection for Machine Condition Monitoring: A Domain Generalization Baseline, *2023 31st European Signal Processing Conference (EUSIPCO)*, Helsinki, Finland, 2023, pp. 191–195. https://doi.org/10.23919/EUSIPCO58844.2023.10289721

[29] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, F. Wei, WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 16, No. 6, pp. 1505–1518, October, 2022. https://doi.org/10.1109/JSTSP.2022.3188113

[30] T. Khandelwal, R. K. Das, E. S. Chng, Sound Event Detection: A Journey Through DCASE Challenge Series, *APSIPA Transactions on Signal and Information Processing*, Vol. 13, No. 1, Article No. e3, February, 2024. http://dx.doi.org/10.1561/116.00000051

Biographies

Jian Jun (JJ) Zeng received his M.S in Electronic and Communication Engineering from Beijing Jiaotong University, in 2013. He is currently PhD candidate in electronic information at Tianjin University. He is also an entrepreneur in intelligent vehicle cybersecurity domain with AI and big data technology, he is CEO and created INCHTEK.AI from scratch at Nov 2020, received more than 20 OEMs as customer, including BMW, FAW, DFAC, GAC, BAIC etc.

Jianguo Wei is currently a Professor at the College of Intelligence and Computing in Tianjin University. He received his M.S. degree Tianjin University, China, and Ph.D. degree in the Japan Advanced Institute of Science and Technology in 2004 and 2007, respectively. His research interests include articulatory modeling, modeling coarticulation, articulatory-based speech synthesis, and image processing for articulatory analysis.