Sound Event Detection Via Pervasive Devices for Mobility Surveillance in Smart Cities

Matteo Sammarco and Trevor Zeffiro
Stellantis
Email: {matteo.sammarco,trevor.zeffiro}@stellantis.com

Luana Gantert and Miguel Elias M. Campista
Universidade Federal do Rio de Janeiro (UFRJ)
GTA/PEE-COPPE/DEL-POLI, Brazil
Email: {gantert,miguel}@gta.ufrj.br

2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) | 979-8-3503-0436-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/PerComWorkshops59983.2024.10503381

Abstract—**Smart cities and Intelligent Transportation Systems rely upon the deployment of sensors in strategic areas for such purposes as crime prevention, urban planning, and road safety. In this paper, we rely on the pervasiveness of smartphones and microphones inside moving vehicles to propose a sound-based event detection system which does not require static sensing infrastructure. We train an embedded Deep Neural Network model able to identify potentially dangerous events like car accidents or emergency vehicles approaching from recorded sounds. We evaluate our model on a large novel dataset of sounds recorded inside the car cabin with audio data augmentation techniques applied thereon. We further evaluate model performance after model quantization, or the addition of environmental noise. Results show an excellent detection accuracy for dangerous events achieving a Matthews Correlation Coefficient (MCC) of 0.95.**

I. Introduction

The growth of vehicular communications worldwide pushes the creation of monitored urban areas, called smart cities, through participatory, or pervasive sensing. Smart cities aim to increase road safety, prevent crimes, provide better urban planning, and manage traffic by analyzing data usually coming from sensing infrastructures composing large surveillance systems [1].

Typical road surveillance systems rely on a coordinated set of cameras [2] or microphone arrays [3], [4]. This infrastructure becomes expensive in both a computationally and monetarily, and does not scale well in large cities. Also, static infrastructure may not follow the real traffic density and have reduced performance due to environmental conditions such as fog, rain, or background noise. Thus, instead of cameras or microphones, an alternative for surveillance systems is to rely on driver participation. Waze is an example of a successful smartphone app for road navigation that leverages crowdsourced reports sharing warnings and other helpful information among drivers. Nevertheless, it requires human interaction, which may introduce bias, reporting time delay, lower reliability, and coverage issues [5]. Additionally, the use of smartphones during driving is a source of hazards itself.

This paper proposes a dataset and a model for audio event recognition from inside the car cabins to further create a participatory road surveillance system for mobility safety. The goal is to recognize crashes or other hazardous situations without static sensing infrastructure or driver interaction.

Audio streams are recorded and analyzed by drivers' smartphones or by the in-vehicle infotainment (IVI) system within the moving vehicles. Modern smartphones and IVIs have both microphones and enough computational power for signal analysis and processing. Moreover, relying on a audio analysis has a twofold advantage: the pervasiveness of devices embedding a microphone (smartphones), and avoiding rigid device positioning, which would be the case with cameras.

We assume a standard five-layer system for our audio surveillance system composed from the bottom up of data acquisition, background subtraction, event classification, event positioning and tracking, and situation analysis. In this paper, however, we focus on the third layer for event classification, which aims to discern Events of Interest (EoI) among a large sequence of events and signals. In particular, we want to distinguish events dangerous to drivers' safety (e.g., car accidents, slippery pavement, and emergency vehicle approaching at high speed) from the sounds of normal driving activity. To accomplish this task, the audio stream is split into small segments, and a spectrogram image is derived from each one of them. Then, a DNN (Deep Neural Network) is trained to classify such spectrogram images. As DNN, we employ a MobileNetV2 model [6] which presents a good trade-off between classification accuracy and inference speed in mobile applications. Nevertheless, DNN models do not generalize well if not trained on a very large and appropriate dataset. Therefore, we have preliminary created a novel dataset of audio signals recorded from inside the car cabin. Furthermore, we extended such dataset through signal manipulation and new signals generation with the use of a second DNN. We evaluate the audio event detection model with standard machine learning metrics with and without extra background noise and with or without optimizations for low-capability devices (IVIs or older smartphones). We show that our system can correctly detect car crashes, car horns, tire skidding, sirens, and other sounds listened inside the cabin.

In a nutshell, our main contributions are:

- We create a large dataset of driving and road environmental sounds recorded within vehicles. As such dataset presents unique aspects, we share it with the research community [1].
- We build a sound classification mobile application embedding a DNN model and we experiment its perfor-

[1] Scripts to recreate the dataset are available at https://github.com/Githeo/NINA-Dataset.

mance in various scenarios: presence of different types and levels of noise and post-training model optimization. [2]. Such experiments evaluate its pervasiveness on limited capacity devices.

This paper is structured as follows: Section II details the dataset created and the data augmentation techniques involved. Section III presents the model for audio classification as well as the classification results with and without post-training optimization. We evaluate the classification performance in the presence of colored noise in Section IV. We examine the system effectiveness and adoption aspects in Section V, while Section VI deals with related work. Finally, Section VII concludes this paper and draws future directions.

II. NINA Dataset

Although some urban and road sound collections exist [4], [7], [8], [9], [10], [11], none of them record audio events from inside the car cabin. It is extremely costly to reproduce some classes of sound (e.g., impacts among real vehicles) and very difficult to artificially create them. In addition, we need to distinguish the sounds of potential hazards from all the other sounds linked to normal driving activity. Hence, we have decided to create a novel dataset of urban and road-related sounds recorded inside the car cabin, called the NINA (Naturalistic IN-vehicle Audio) dataset. We share this dataset with the research community for other pervasive applications. In particular, we took advantage of the Audacity audio editing software [12] to annotate audio tracks recorded by dashcams and published on YouTube. We focus our labeling task on the following seven classes grouped in two areas:

- **Events of Interest (EoI)**. These classes generate alerts to share with other drivers:
  - **Crash**. Accidents between any kind of vehicle and with a different degree of severity.
  - **Tire**. Tire skidding sounds, often provoked by a harsh braking to avoid an impact or by slippery pavement.
  - **Emergency**. Siren sounds produced by EVs (Emergency Vehicles). Records come from police cars, ambulances, and fire trucks in Italy, France, Germany, and Netherlands.
  - **Horn**. Various kinds of car horn.
- **Internal**. Sounds generated inside the ego-vehicle and related to a normal driving behavior:
  - **Driving**. This class includes the sound of the vehicle engine running at constant speed as a normal driving activity.
  - **Voice**. People talking inside the vehicle in different languages.
  - **Music**. Radio music clips of various kind.

A further advantage using the NINA dataset is that audio clips are genuine: they include noise and background sounds that we could expect in real life (e.g., windscreen wiper swinging, turn signals, strong wind, etc.). Also, since the recording microphones are different too, we are close to real in field conditions.

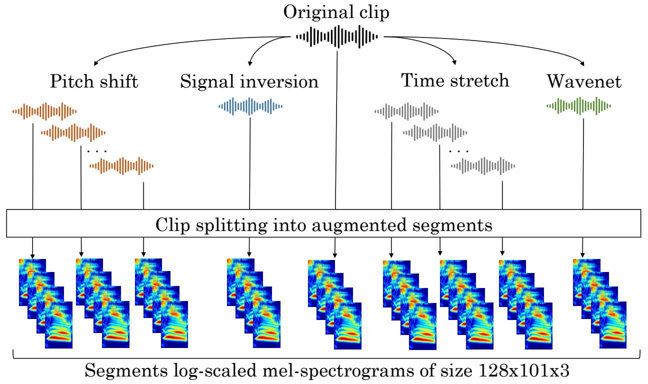

Original clip
Pitch shift
Signal inversion
Time stretch
Wavenet
. . .
. . .
Clip splitting into augmented segments
Segments log-scaled mel-spectrograms of size 128x101x3

Fig. 1. Original audio clip augmentation and splitting into fixed size spectrograms.

All the audio clips are resampled to 22,050 Hz and normalized in amplitude. They all have a different duration: some clips could be very short, e.g., short car horn activation or fast tire skidding. To catch such sounds, we split clips in segments of 12,800 samples (about 0.5 seconds at 22,050 Hz) with a 50% overlap to grasp sounds falling at the middle of two consecutive segments. Then, for each segment we compute the log-scaled mel-spectrogram with 128 bands, FFT window and hop size of 1024 and 128 samples, respectively. Finally, we compute first and second order deltas (with the default python library *librosa* settings), resulting in segments $\in \mathbb{R}^{128x101x3}$. Ultimately, a serialized version of the dataset has a size of 15 GByte. Hereafter, we consider segment spectrogram images in place of raw audio signals.

A. Data augmentation

CNNs (Convolutional Neural Networks), as DNNs in general, show an effective ability in classification tasks, but they need a large amount of training data in order to well generalize and classify unseen data. We apply the following augmentation techniques on the raw audio signals:

- **Pitch shifting**. This manipulation shifts a signal up or down in frequency keeping the same time duration. Each original audio clip is pitch shifted by semitones in the set {-4, -3, -2, -1, 1, 2, 3, 4}.
- **Time stretching**. Opposite to the pitch shift, this effect changes the speed of an audio signal without affecting its pitch. We use 6 time stretch scaling values: {0.7, 0.8, 0.9, 1.2, 1.5, 2.0}.
- **Signal inversion**. The audio signal is reversed while sound classes remain unchanged to human listeners. Even people talking clips remain distinguishable although hardly intelligible. It is equivalent to flip horizontally the spectrogram image.
- **WaveNet generated signals**. WaveNet is a DNN for generating raw audio signals initially developed for TTS (Text-To-Speech) applications [13]. WaveNet directly

[2] A demo is available at https://github.com/Githeo/CarAudioSensing-Demo.

TABLE I
AMOUNT OF ORIGINAL CLIPS AND RELATED AUGMENTED SEGMENTS.

Type	**Class**	**Original clips**	**Augmented segments**
EoI	Crash	751	32368
	Tire	186	7565
	Horn	261	11968
	Emergency	259	123828
Internal	Driving	295	56168
	Voice	422	34170
	Music	198	43027
Sum		2372	309094

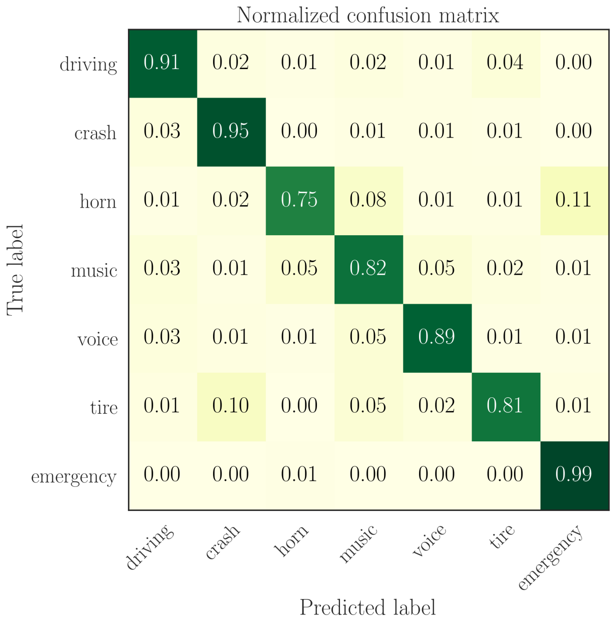


Fig. 2. Multi-class normalized confusion matrix for test set classification (values rounded at the second decimal). $MCC = 0.905$.

generates samples of a raw audio waveform where each sample probability is conditioned on the samples at all previous timestamps. In order to reduce computational cost with respect to recurrent neural networks (RNN), WaveNet uses dilated convolutions to keep a low number of hidden layers and a μ-law companding transformation to reduce the softmax layer from 2^{16} to 2^8 possible output values. WaveNet eventually outputs new audio signals similar to the original ones but with a certain mutation degree. The new signals have a resolution of 16 KHz and are re-quantized at 16 bits.

Figure 1 illustrates the data augmentation and splitting applied to the original audio clips, while Table I summarizes the amount of original clips and augmented segments per class. Original clips have all different duration in according of their nature: crash and tire skidding sounds usually last few seconds, whereas music and driving clips have longer duration. For this reason, even if original driving or music clips are less numerous than crash clips, they are split into more numerous augmented segments. For each original clip, an equal group id is assigned to its segments and augmented segments. The idea of assigning a common id is to prevent segments belonging to the same root clip presenting in multiple different subsets during the training, validation, and test dataset split. A 5% of non-augmented segments, including all the classes, are randomly chosen to compose the test set used to evaluate the model generalization at the very end. All the other segments having the same group id are completely discarded. The network weights used to classify the test data result in the best accuracy on the validation set (15% of the remaining dataset) after a 10-fold cross validation training.

III. MODEL AND SOUND CLASSIFICATION RESULTS

One of the most popular approaches for sound classification is to consider the spectrogram image of the sounds and then leverage a deep convolutional neural network to classify images [14], [15], [16], [17], [18], [19]. In this work, we follow this stream, mainly for training computational cost reasons with respect to analyzing the raw audio signal with recurrent neural networks [20], [21] and for inference performance with respect to feature-oriented models [22], [23], [24], [25]. Moreover, the model we use, MobileNetV2, is designed to run on the edge for mobile visual recognition tasks including classification, object detection, and semantic segmentation [6]. Its architecture is especially designed to efficiently run on devices with low computational power (drivers' smartphone or IVIs in our case) due to its optimal trade-off between accuracy and model size or complexity [26].

MobileNetV2 uses depthwise separable convolution as efficient building blocks. Also, it removes non-linearities in the narrow layers and shortcuts connections between them in order to maintain representational power. For transfer learning purposes, we load the MobileNetV2 architecture with weights inherited from ImageNet and we add a dense layer of 256 units and ReLU activation followed by a 7 units dense layer with softmax activation. We took advantage of `Tensorflow` with `Keras` as high level neural-network libraries for model development and training.

A. Classification result

Figure 2 shows the normalized confusion matrix for the test set classification. Classification is very close to perfection for two very important classes, crash and emergency, with 95% and 99% respectively. Also, the event-free driving is well recognized: 91%. Car horn sounds, instead, are often misclassified as emergency sirens. Actually, they present some similarities: a high amplitude, sustained for long time at some specific frequencies (fundamental and harmonics). As an image then, they both show long horizontal peaks of amplitude at some frequencies. Music can contain any instrument, voice and sustained rhythm, thus 10% of their samples are misclassified as horn or voice samples. Finally, 81% of tire skidding sounds are well classified.

As the NINA dataset is unbalanced, to appreciate the final test set classification performance, we also rely on the Matthews Correlation Coefficient (MCC), which is equal to 0.905, very close to the state of the art on image classification results.

B. Post-training optimization

The model for sound recognition is intended to be embedded and run in pervasive devices like drivers' smartphone, IVIs or any other ad hoc hardware. Therefore, we further experiment performances after model optimization for limited hardware capabilities both for storage and inference execution latency. In particular, we employ post-training quantization techniques which, as a first step, can reduce the model size, and then improve CPU and hardware latency in exchange for a degradation in accuracy.

The model trained in the previous section initially has a 30 MByte size on disk. Taking advantage of the TensorFlow Lite conversion tool, the model is compressed to 10 MByte. At this point, we test two quantization methods:

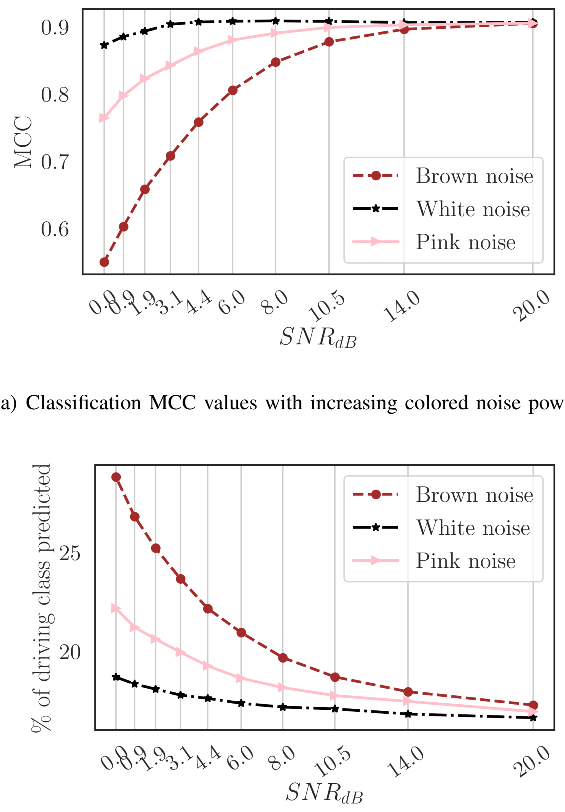
- **Float16 quantization**. Model parameters are quantized to a 16-bit floating point representation which halves both the overall size of the model and the CPU bandwidth used during the parameters loading process. Some hardware is also optimized to work with 16-bit floating points numbers resulting in reduced latency.
- **Hybrid quantization**. Weights are converted from 32-bit floating point numbers to the 8-bit integers, while keeping biases and activations with the original representation. Also computation is mixed between floating point and integer. Benefits include a model size reduced 4 times and an inference speed 2 times faster.

In the first case, model size is equal to $5\,\mathrm{Mbytes}$, while it executes in 32-bit floating point numbers anyway, thus its accuracy remains mostly unchanged: $MCC = 0.904$.

With the hybrid optimization, weights are converted from 8-bits of precision to floating point at inference time. This conversion is done once and cached to reduce latency. Nevertheless, classification results highly degraded ($MCC = 0.72$), even though the model size is half of the precedent optimization. Considering this last result, we renounce to further convert the model to a full integer quantization of weights and activations.

IV. Noise robustness

Sensing systems, especially audio and video-based, often deal with the problem of background subtraction. The aim is to reduce the noise affecting and degrading images and signals. However, trying to clean a signal and run inference continuously can become computationally expensive and time demanding. On the other hand, the NINA dataset contains audio clips recorded *in vivo*, thus already including other road background sounds like city bustle, traffic noise, heavy rain, or strong wind. The classification presented in Section III-A implicitly takes into account this aspect.



(a) Classification MCC values with increasing colored noise power.

(b) Percentage of samples classified as "driving" class.

Fig. 3. Classification results adding white, pink and brown noise to the test set samples.

Nevertheless, in this section, we experiment and estimate the impact of noise generated by the recording equipment or by the vehicle itself with its air conditioning vents or engine. In fact, microphones are prone to add self-noise due to high temperatures (called thermal noise), current running in the circuitry (called shot noise) and subsequent amplification. We model such kinds of noise adding a white Gaussian noise to the test set samples. Moreover, dashcams and smartphones are usually placed on the front windscreen, or immediately up to the vehicle dashboard. In such conditions, the microphones are closer to the vehicle air conditioning vents and to the engine. We model these cases adding to the testing samples pink and brown noise signals, respectively, as they present a spectral density more concentrated to low frequencies.

Figure 3(a) shows the classification MCC values when adding colored noise. As shown on the x-axis, noise signals never overcome in power the original sounds, but they span from having the same power ($SNR_{dB} = 0$) till a $SNR_{dB} = 20$. Brown noise has the most notable impact, as shown through sharpest decrease in MCC with decreasing SNR. Also the pink noise has a quite important degrading effect on classification for low SNR values, while the results with white noise remain acceptable (MCC = 0.87 as minimum). Figure 3(b) shows the reason for the MCC decreasing: most of the misclassification is related to sounds classified as "driving" while being actually another class. The driving class, in fact, does not include any special event, but just the running engine at different speeds, which sounds like a brown noise. The white noise instead masks events, making the spectrogram more blurry. These results are obtained with a non-quantized model.

V. Discussion

In this section, we examine concerns impacting the adoption and the effectiveness of the proposed pervasive sensing system.

Privacy. The utilization of a microphone often raises concerns about eavesdropping. It is worth to note, that the model presented in Section III is designed and implemented to run embedded in a mobile application and it operates the classification on the fly. Thus, what is listened within the vehicle, remains inside the vehicle. Also, GPS locations are transmitted only when an EoI occurs.

Coverage and adoption. The territory coverage of a participatory system is proportional to the number of users involved. Its initial adoption follows the same rule: more users involved, more EoI labeled and even more users are attracted. Conversely, if only few users contribute, new users will not be encouraged to participate unless other incentives are given [27]. Such impasse is overcome if public transports are involved. For instance, the sensing vehicles include all the public buses, the surveillance system will cover the main city road all day long.

Complementarity. The usage of commodity and pervasive devices and technology (namely smartphones and Wi-Fi) makes the system easy to interact with pre-existing surveillance road side units (RSU) and infrastructure. Thus, the proposed participatory approach can coexist with other surveillance systems already deployed in smart cities.

VI. Related Work

A review about audio surveillance is proposed in [3]. Foggia et al. are among the first authors designing a road surveillance system via the deployment of a road side large microphone array [28], [29], [24], [4], [30], [31]. Initially, their audio classification was based on a combination of bag of words and SVM with features extracted from the raw audio signals. Successively, DNN models, in the form of CNN, have proved their effectiveness in classifying audio signals from their spectrogram representation. SoReNet [32] is just an example of CNN-based audio surveillance system among others [33], [34], [35], [36], [37]. The following two sets of problems are neglected in such works. Besides model performance, the practical deployment of these systems is directly tied to their scalability, cost, and modularity. Hence, relying on a fixed infrastructure and on computationally expensive DNN models is not an approach ready for wide adoption. The second issue regards the training of deep learning models: they require a massive amount of data in order to well generalize on unseen signals and none of the mentioned works perform audio data augmentation beforehand.

Specific works on sound classification focus on audio data augmentation too [15], [17], [38], [39], [40], [41]. Some techniques introduced in these works, like time stretching, pitch shifting, and noise injection, are widely adopted. Other approaches like dynamic range compression, random cropping, and equalization are more linked to specific domains (e.g., techniques for speech recognition problems). Our work also relies on these known standards in addition to signal inversion and transfer learning using WaveNet DNN, ensuring that the signal semantic would not change.

VII. Conclusion

In this paper, we have presented and experimented a sound event detection model for mobility safety. It is the core of a participatory sensing system for road surveillance and mobility safety for smart cities and ITS. Large-scale participation depends upon the ability to run such model on pervasive devices with an embedded microphone (e.g., drivers' smartphones and IVIs). For this reason, we evaluated the model performance with the presence of different kinds and level of noise, and reducing the model size and inference time devices with limited capacity.

EoI are recognized via a CNN model specifically trained to discern sounds by their spectrogram image representation. Trained and tested on our novel, large, dataset of sounds recorded from inside the vehicle cabin, the model has a MCC = 0.905 over seven classes.

We plan to add other EoI sounds in the dataset like gun shots, people screaming, or pothole. Such audio signals still must be recorded within the vehicles. As a future work, we would like to include event localization using one, two, or more than two sensing vehicles.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. This paper was also funded by CNPq, CAPES, FAPERJ, FAPESP (2015/24485-9, and 2014/50937-1), and the Development and Research Foundation – Fundep – Rota 2030.

References

[1] H. Arasteh, V. Hosseinnezhad, V. Loia, A. Tommasetti, O. Troisi, M. Shafie-khah, and P. Siano, "Iot-based smart cities: a survey," in *2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC)*. Florence, Italy: IEEE, Jun. 2016, pp. 1–6.

[2] L. Tian, H. Wang, Y. Zhou, and C. Peng, "Video big data in smart city: Background construction and optimization for surveillance video processing," *Future Generation Computer Systems*, vol. 86, pp. 1371–1382, 2018.

[3] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, Feb. 2016. [Online]. Available: https://doi.org/10.1145/2871183

[4] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio Surveillance of Roads: A System for Detecting Anomalous Sounds," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, Jan. 2016.

[5] M. Amin-Naseri, P. Chakraborty, A. Sharma, S. B. Gilbert, and M. Hong, "Evaluating the reliability, coverage, and added value of crowdsourced traffic incident reports from waze," *Transportation Research Record*, vol. 2672, no. 43, pp. 34–43, 2018. [Online]. Available: https://doi.org/10.1177/0361198118790619

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA: IEEE, Jun. 2018, pp. 4510–4520. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00474

[7] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 1015–1018. [Online]. Available: http://doi.acm.org/10.1145/2733373.2806390

[8] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. Budapest, Hungary: European Association for Signal Processing (EURASIP), Aug. 2016, pp. 1128–1132.

[9] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[10] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, USA: IEEE, March 2017, pp. 776–780.

[11] S. Adavanne, A. Politis, and T. Virtanen, "TAU Spatial Sound Events 2019 - Ambisonic and Microphone Array, Development Datasets," Feb. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2599196

[12] A. Team, "Audacity Software," https://audacityteam.org/, 2019, accessed: 2019-09-30.

[13] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016.

[14] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Boston, USA: IEEE, Sep. 2015, pp. 1–6.

[15] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, March 2017.

[16] J. Abeßer, S. Mimilakis, R. Gräfe, and H. Lukashevich, "Acoustic scene classification by combining autoencoder-based dimensionality reduction and convolutional neural networks," in *Detection and Classification of Acoustic Scenes and Events Workshop, DCASE 2017*. Munich, Germany: Tampere University of Technology, 10 2017, pp. 7–11.

[17] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, "Deep convolutional neural networks and data augmentation for acoustic event recognition," in *Interspeech 2016*. San Francisco, USA: International Speech Communication Association (ISCA), 2016, pp. 2982–2986. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-805

[18] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, and et al., "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, USA: IEEE, Mar. 2017, pp. 131–135. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2017.7952132

[19] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification and domestic audio tagging," DCASE2016 Challenge, Tech. Rep., September 2016.

[20] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Lujiazui, China: IEEE, 2016, pp. 6440–6444.

[21] L. Müller and M. Marti, "Bird sound classification using a bidirectional LSTM," in *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*. Avignon, France: Springer Nature Switzerland, Sep. 2018, pp. 1–13. [Online]. Available: http://ceur-ws.org/Vol-2125/paper_134.pdf

[22] G. Takahashi, T. Yamada, S. Makino, and N. Ono, "Acoustic scene classification using deep neural network and frame-concatenated acoustic feature," DCASE2016 Challenge, Tech. Rep., September 2016.

[23] M. Sammarco and M. Detyniecki, "Crashzam: Sound-based car crash detection," in *Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems, VEHITS 2018*. Funchal, Madeira, Portugal: SciTePress, Mar. 2018, pp. 27–35. [Online]. Available: https://doi.org/10.5220/0006629200270035

[24] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento, "An ensemble of rejecting classifiers for anomaly detection of audio events," in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*. Beijing, China: IEEE, Sep. 2012, pp. 76–81.

[25] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," in *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2013*. Krakow, Poland: IEEE, Oct. 2013, pp. 81–86.

[26] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, p. 64270–64277, 2018. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2018.2877890

[27] F. Restuccia, S. K. Das, and J. Payton, "Incentive mechanisms for participatory sensing: Survey and research challenges," *ACM Trans. Sen. Netw.*, vol. 12, no. 2, Apr. 2016. [Online]. Available: https://doi.org/10.1145/2888398

[28] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, and N. Petkov, "Car crashes detection by audio analysis in crowded roads," in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. Karlsruhe, Germany: IEEE, 2015, pp. 1–6.

[29] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, and I. Fellow, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, no. C, pp. 22–28, 2015.

[30] R. Leiba, F. Ollivier, R. Marchiano, N. Misdariis, J. Marchal, and P. Challande, "Acoustical Classification of the Urban Road Traffic with Large Arrays of Microphones," *Acta Acustica united with Acustica*, vol. 105, no. 6, pp. 1067–1077, 2019. [Online]. Available: https://hal.sorbonne-universite.fr/hal-02457922

[31] S. Jennings and J. Kennedy, "Acoustical Classification of the Urban Road Traffic with Large Arrays of Microphones," *Acta Acustica united with Acustica*, vol. 105, no. 6, pp. 1042–1052, 2019.

[32] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "SoReNet: a novel deep network for audio surveillance applications," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. Bari, Italy: IEEE, 2019, pp. 546–551.

[33] P. Foggia, A. Saggese, N. Strisciuglio, M. Vento, and V. Vigilante, "Detecting sounds of interest in roads with deep networks," in *2019 20th International Conference on Image Analysis and Processing (ICIAP)*. Trento, Italy: Springer Nature Switzerland, Sep. 2019, pp. 583–592.

[34] V. Morfi and D. Stowell, "Deep learning for audio event detection and tagging on low-resource datasets," *Applied Sciences*, vol. 8, no. 8, p. 1397, 2018.

[35] A. Roberto, A. Saggese, and M. Vento, "A deep convolutionary network for automatic detection of audio events," in *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, ser. APPIS 2020. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3378184.3378186

[36] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, 2016, pp. 6460–6464.

[37] Z. Huang, C. Liu, H. Fei, W. Li, J. Yu, and Y. Cao, "Urban sound classification based on 2-order dense convolutional network using dual features," *Applied Acoustics*, vol. 164, p. 107243, 2020.

[38] L. Nanni, G. Maguolo, and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecological Informatics*, vol. 57, p. 101084, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574954120300340

[39] P. Vinayavekhin, S. Wang, D. Wood, and R. Tachibana, "Shuffling and Mixing Data Augmentation for Environmental Sound Classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York, USA: DCASE Community, Oct. 2019, pp. 109–113.

[40] Y. Chen and H. Jin, "Rare Sound Event Detection Using Deep Learning and Data Augmentation," in *Interspeech 2019*. Graz, Austria: International Speech Communication Association (ISCA), 2019, pp. 619–623.

[41] V.-v. Eklund, "Data Augmentation Techniques for Robust Audio Analysis," Ph.D. dissertation, Tampere University, Faculty of Information Technology and Communication Sciences, 2019.