Enhanced railway sound event detection using YAMNet and classification criteria

Erwann Betton-Ployon, Abbes Kacem, Jerome I. Mars, Nadine Martin

25-29 August 2024
inter.noise
NANTES FRANCE

Enhanced railway sound event detection using YAMNet and classification criteria

Erwann Betton-Ployon[1]
ACOUSTB
24 Rue Joseph Fourier, 38400 Saint-Martin d'Hères - France

Abbes Kacem[2]
ACOUSTB
24 Rue Joseph Fourier, 38400 Saint-Martin d'Hères - France

Jérôme Mars[3]
GIPSA-Lab
Universite Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-Lab, 38000 Grenoble - France

Nadine Martin[4]
ASTRIIS
710 Avenue de la Motte-Servolex, 73000 Chambéry, France - France

ABSTRACT

The segmentation-by-classification method has become a popular way to detect sound events. It uses neural networks, trained to detect sound sources on short signals (one tenth of a second to a second), or long-term signals divided into a succession of short segments. In this paper, we focus on detecting train pass-bys from long term signals, where it is assumed that railway noise is higher than ambient noise. Using the neural network YAMNet, we show that a processing stage is necessary to improve the segmentation of such events and reduce the high false positive rate. The applied criteria are related to the nature of a train pass-by, lasting several seconds and with broad frequency band. Around 90% of events are detected with proper boundaries. However, we observe that YAMNet is not designed to distinguish railway vehicles from other vehicles. False positive rate remains high whenever other vehicles travel near the measurement location. Additional AI models are proposed to perform this specific distinction. Their efficiency depends on train type, as recent passenger train noise resembles roadway noise. Otherwise, false positive rate decreases below 10%, and the railway noise contribution estimate aligns with the reference level, within a 0.5 dB(A) margin on average.

[1]erwann.betton-ployon@egis-group.com

[2]abbes.kacem@egis-group.com

[3]jerome.mars@gipsa-lab.grenoble-inp.fr

[4]nadine.martin@astriis.com

1. INTRODUCTION

Noise pollution has become an increasing matter of concern for the last decades. Several studies linked up high noise exposure with cardiovascular problems and blood pressure increase, for workers [1] and children [2, 3]. Traffic noise is among the sources responsible for the higher noise levels, when averaged over time [4]. Often, noise standards are applied on a national or continental scale, and distinguish roadway, airborne and railway noise. Thus, we focus on measuring railway noise, in residential areas that are located near railway lines.

According to French standards [5], railway noise must be evaluated over long periods, from one day up to one week, to make sure measurements reflect the actual situation. On these long-term acoustic signals, train pass-bys are identified and used to estimate railway contribution to overall noise through several specific indicators. Our aim is to automate the process of detecting train pass-bys on acoustic signals, while ensuring the estimated indicator values are comparable to those obtained through manual identification.

Sound Event Detection (SED) has been explored in many research efforts, especially since the rise of deep learning techniques and neural networks [6, 7]. Some authors [8] and [9] used machine learning and statistical criteria applied on the evolution of noise levels to detect train pass-bys. However, limits were observed for generalisation to new acoustic environments.

Non-linear models like neural networks showed promising results on SED [6, 7]. To our knowledge, neural networks have not been used to solve a train pass-by detection task. Still, [10] worked on detecting abnormal events in noisy embedded railway environments, using convolutional recurrent neural networks. The high obtained accuracy shows the ability of convolutional neural networks (CNN) combined with spectrograms to separate sound sources in railway environments. Other works also used mel-spectrograms to solve SED tasks [6, 7].

However, in outdoor acoustic environments, the spectrum of possible sound sources is usually broader [6]. The risk of interfering noise is also higher, and a complete, diversified database is often required to obtain an accurate classifying neural network. The cost of collecting and labelling acoustic data at high scale makes it difficult to only use an internal database to train an environmental sound classifier. Therefore, we choose to rely on the external classifier YAMNet, and use our own expertise and dataset to process and narrow its detection results.

2. PROPOSED TRAIN PASS-BYS DETECTION METHOD

Regarding this matter, a staged approached is chosen, represented in Figure 1. To assess complex sound environments and the variety of sources, a pre-classification is made. This first step is called "YAMNet `Vehicle` Detection" (YVD) and will be detailed in Section 4. An "Event Recomposition and Filtering" (ERF) process follows, to rectify misclassified frames and overlapping sources issues (Section 5). A final classifier (Section 6), trained on our data, is used to differentiate train pass-bys from other sound sources.

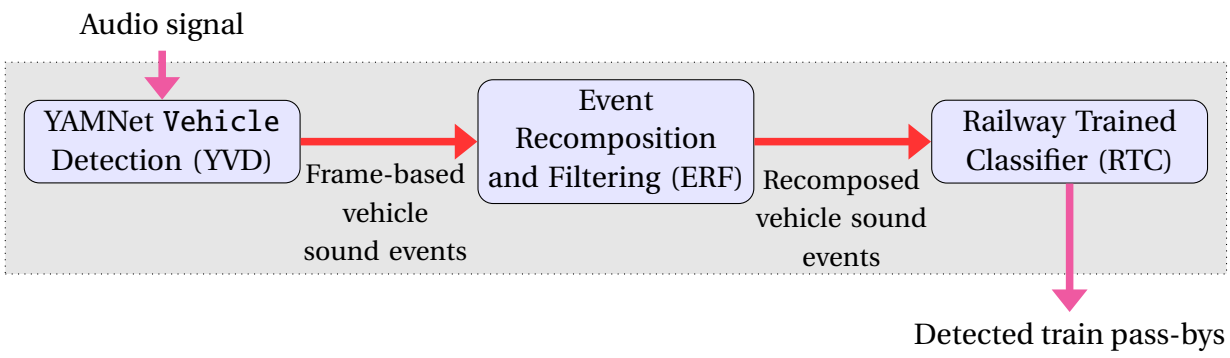

Figure 1: Description of the proposed train pass-by detection method.

3. MEASUREMENT CONDITIONS AND EVALUATION

After each stage, all detected events will be evaluated on acoustic measurements, recorded in real environments from 17 different locations. This section details the measurements conditions and our evaluation method.

Each record lasts 24 hours, sampled at 25600 Hz on a single 32-bits channel. The objective of such measurements is to evaluate railway noise contribution in living areas. Thus, sound level meters are placed near railway lines (mostly between 5 and 15 m away, always less than 50 m).

From the combined 408 hours of measurements, one can distinguish 3 contexts that give an indication on the expected sound environment. During the evaluation process, the "rural without road traffic", "rural with road traffic" and "urban" contexts are separated. For further use, such distinction helps anticipating the method accuracy on a long-term measurement, compared to SNR (Signal-to-Noise Ratio), which varies with each train pass-by.

At each stage of the method (Figure 1), the precision, recall and F_1 (Eq. 1) are measured.

$$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad ; \quad \text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad ; \quad \text{F}_1 = \frac{2.\text{precision}.\text{recall}}{\text{precision}+\text{recall}} \tag{1}$$

where TP stands for true positive, FP for false positive and FN for false negative. A train pass-by is considered to be detected (True Positive) if most of its corresponding frames are classified as train. In other words, a 10-second long pass-by is detected if more than 5 seconds are classified as train over the whole event. Also, a longer-than-expected detected event will count as one true positive: a 30-second long detected event will count as a true positive if it includes any real train pass-by.

4. YAMNET VEHICLE DETECTION (YVD)

As a first step, a broad classifier is used on the signal, formerly subdivided into frames of the same duration. This classification aims at filtering out time spans without any railway related sounds. To do so, the classifier, a neural network, relies on signal transformations to recognize patterns similar to those computed on a training dataset. Mel-spectrograms are an efficient way to represent the signal and highlight critical patterns [6, 7]. Figure 2 shows a typical use of a CNN to classify and detect any occurrence of a specific sound source. The input of a CNN-classifier is an image or a series of images. Each image classification results in a probability vector: a column vector with an occurrence probability for each class of the training dataset.

Regarding the training data, our internal database mainly contains railway related data ; it lacks other environmental sound sources. A fully-trained classifier is then necessary to process the complete environmental acoustic signal. *YAMNet* is a CNN developed and trained by Google on a part of the *Audio Set* dataset. *Audio Set* is a large dataset, containing around 1.8M of labelled segments distributed between 600 classes [11]. 521 of these classes compose the *YAMNet* training dataset. As inputs, the model receives mel-spectrograms: a classic spectrogram on which 64 mel filters are applied, from 125 Hz to 8 kHz. *YAMNet* uses 960 ms long frames with a 50% overlap, meaning that a signal of T seconds duration results in a series of $\lceil \frac{T}{0.96\times 0.5} - 1 \rceil$ probability vectors.

Our use of *YAMNet* corresponds to the process described in Figure 2. At this stage, each probability vector is processed independently. On each vector, the detected class is the one with the highest probability. Among the *Audio Set* 521 classes, 7 of them can relate to a train pass-by, the main ones being `Vehicle`, `Train` and `Rail transport`. Any probability vector having one of these 7 classes as the most likely is kept as a train pass-by.

This reasoning applied to all probability vectors is the outcome of the YVD stage. Figure 3 shows an example of a 30-minutes long measurement in an urban environment, near several roads including a highway further away. The evolution of equivalent sound level, over 1-second periods ($L_{eq,1s}$) is displayed. A dense road traffic on the highway emits sound almost continuously. Besides, 5 events emerge from background traffic noise, marked off by the dashed lines. Red lines are the manually annotated train pass-bys, whereas the other sources are represented by the blue lines. Red-colored areas represent the frames that were classified as train pass-by after YVD.

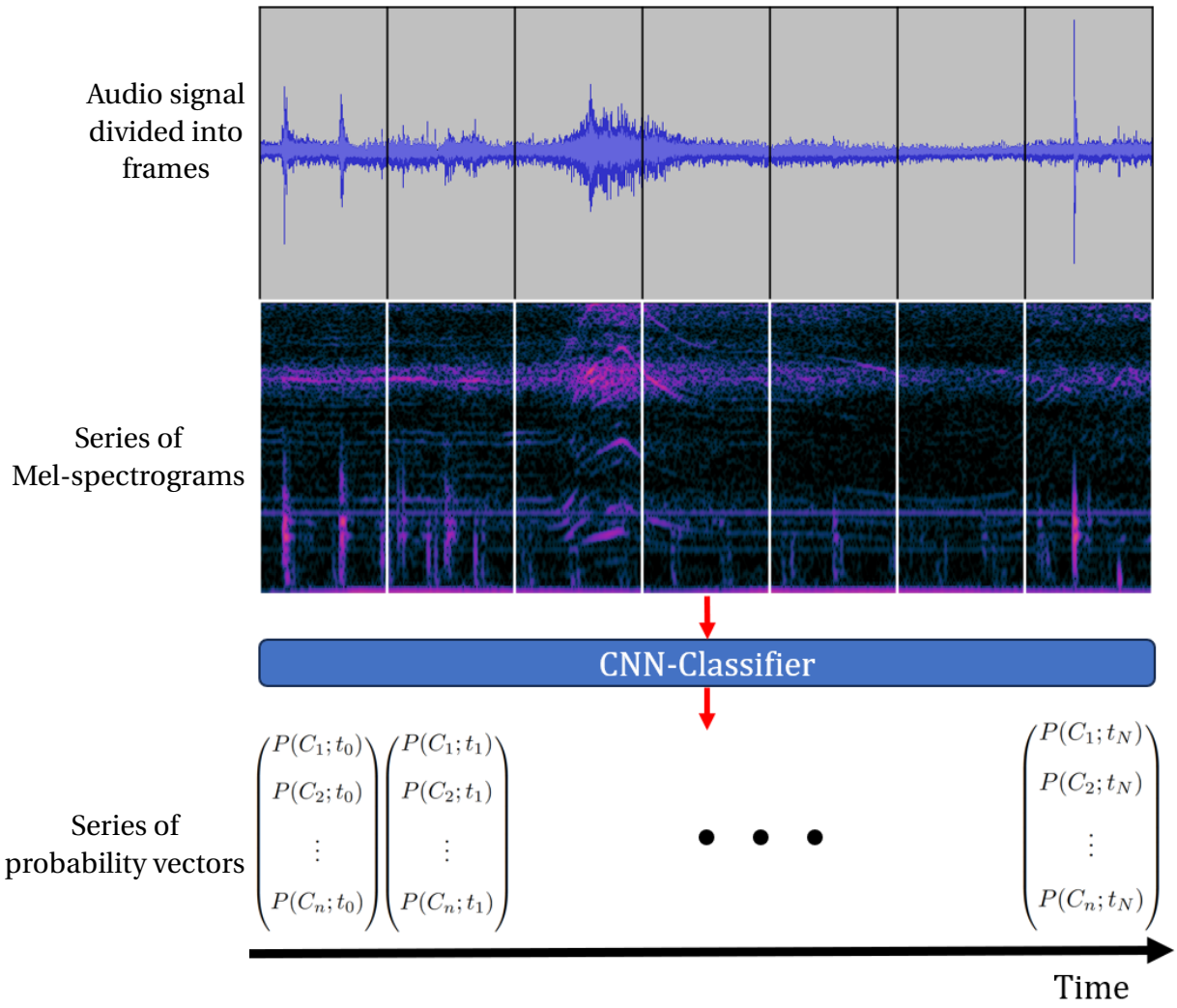

Figure 2: Description of a CNN-based classification use.

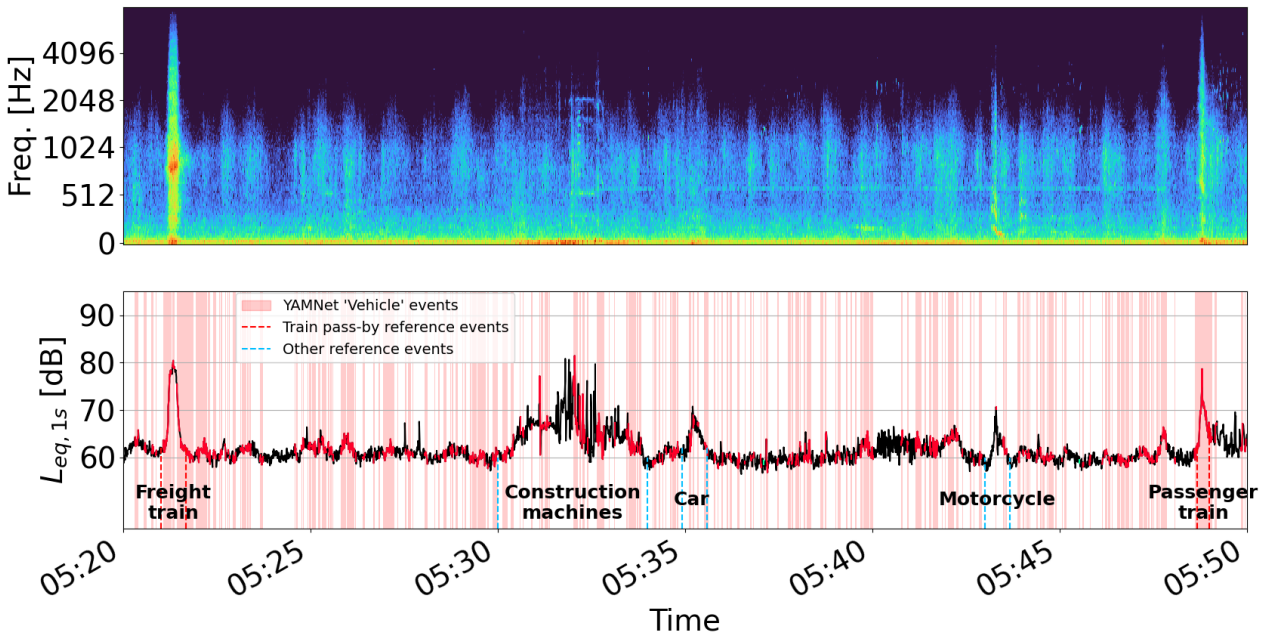

Figure 3: Mel-spectrogram and evolution of equivalent sound level of a 30-minutes long audio measurement in urban context, with the result of the YVD stage.

This outcome reflects how CNNs like *YAMNet* work when applied on long-term signals. Short-term frames are treated independently and longer events present discontinuities, because of a few frames misclassifications. This phenomenon drives up the amount of detected events, far superior than our expectations. Also, among the selected *Audio Set* classes that refer to train pass-by, we observe that the `Vehicle` class is put forward most of the time. Due to *Audio Set* data imbalance and hierarchy between classes [11], *YAMNet* rarely choses precise labels, like `Train` or `Rail transport`, over broad ones. This adds lots of false positives, because of the highway traffic being mixed with train pass-bys in the `Vehicle` class. Table 1 shows the evaluation metrics on the complete dataset, which confirms trends observed in Figure 3.

Table 1: Evaluation of the YVD outcome on the whole dataset, separated in three different measurement contexts.

Context	**♯ of expected pass-bys**	**♯ of detected events**	**Precision (%)**	**Recall (%)**	**F_1 (%)**
Rural w/o road traffic	249	1691	14.7	99.6	25.6
Rural w/ road traffic	287	6392	4.4	99.0	7.7
Urban	893	33916	2.6	99.6	5.1

The 3 chosen metrics highlight different aspects of the YVD outcome. First, the precision is related to the proportion of false positive events (Eq. 1). Most false positives are related to pass-by road traffic noise, so the precision decreases as much as road traffic intensifies. In the 3 environments, the recall is high, almost at 100%: the amount of missed events is very low after the YVD stage. Lastly, the F_1, harmonic mean of precision and recall (Eq. 1), is close to the precision metric. It shows that the area of improvement after YVD is the removal of false positive.

5. EVENT RECOMPOSITION AND FILTERING (ERF)

Considering these intermediate conclusions, the ERF stage (Step 2 from Figure 1) focuses on the two main flaws of the YVD outcome. First, events are recomposed to avoid discontinuities in the complete train pass-by. Then, a second processing stage aims at removing most false positive events, which are known to be mainly caused by road traffic noise.

Regarding the event recomposition, it is assumed that at least one frame of each pass-by is detected in the YVD stage. Otherwise, the train pass-by is completely missed and cannot be retrieved. There exists several ways to define the starting and ending points of a train pass-by. However, a common procedure is to place events boundaries according to the emergence. As soon as the energy level emerges from the background noise, a boundary can be placed. Using this definition of sound events boundaries, a small signal processing technique allowed us to find the emergence instants of each event.

A moving average of the 1-second equivalent sound level evolution ($L_{eq,1s}$) is computed. Then, the detected events are extended to the nearest local minima, in both directions. Empirically, this technique provides a proper event demarcation for train pass-bys. Also, whenever an extended event meets the boundary of another event from the YVD outcome, both are merged into a single larger event.

Even though some false positives are merged, they have not been properly removed yet. To differentiate real train pass-bys from other sound sources, we first deepen the *YAMNet* probability vectors analysis. Railway events can be categorized in the `Vehicle` class, which also contains road

traffic noise. Nevertheless, if some patterns on the spectrogram are related to typical train pass-by noise, the `Train` and `Rail transport` scores also increase. Thus, by applying a threshold on `Train` and `Rail transport` probabilities, some obvious false positives are removed.

Based on the measurement characteristics and train pass-by properties, minimal event duration and emergence are expected. Any event that does not last more than 5 seconds or emerges by more than 5 dB will be filtered out.

ERF stage outcome on the previously-introduced example is shown on Figure 4. From the 171 events proposed after the YVD stage, there are now 6 events in the ERF outcome.

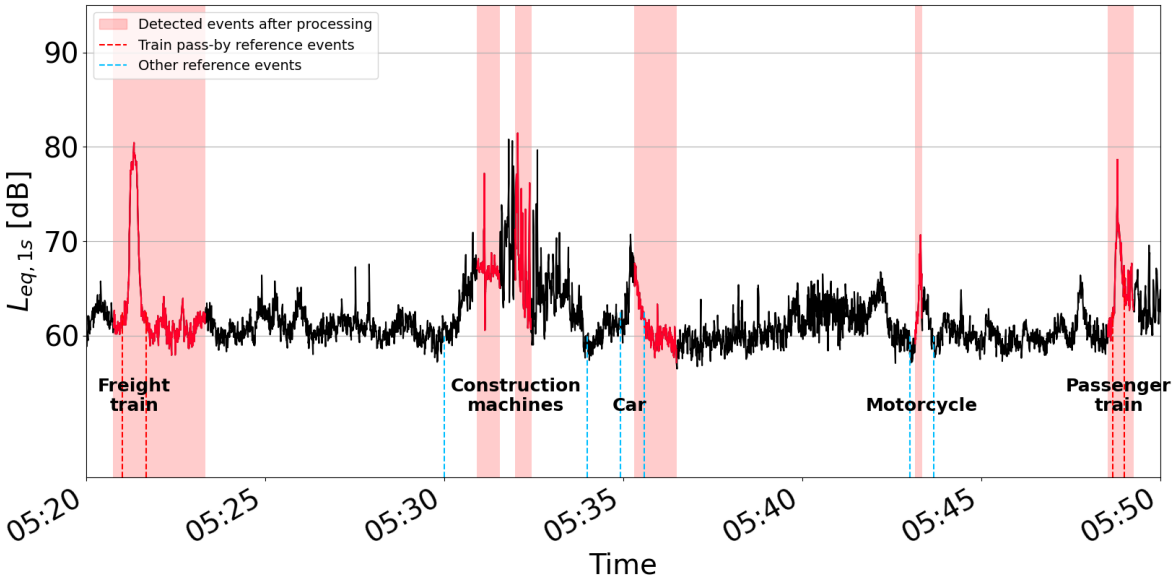

Figure 4: Evolution of equivalent sound level of a 30 minutes long audio measurement in urban context, with a representation of the ERF outcome.

The declining number of proposed events is imputable to both the events expansion and selection criteria. Formerly (Figure 3), the freight train pass-by was divided into 3 different events. These 3 events have been merged at the beginning of the ERF stage, even including some of the surrounding highway traffic noise.

Besides, all the events that were related to the highway traffic are successfully removed. There only remains the emerging vehicles, and a few other interfering sources (construction machines and trucks). Regarding the event demarcation, both trains are fully included in a single event, which corresponds to our expectations, even if the freight train event is wider than the reference boundaries.

Table 2: Evaluation of the ERF outcome on the whole dataset, separated in three different measurement contexts.

Context	**♯ of expected pass-bys**	**♯ of detected events**	**Precision (%)**	**Recall (%)**	**F_1 (%)**
Rural w/o road traffic	249	630	36.0	91.2	51.6
Rural w/ road traffic	287	1555	16.0	86.8	27.0
Urban	893	3747	22.6	94.7	36.5

Over the whole dataset (Table 2), the precision significantly increases, especially in the two contexts with road traffic. More specifically, continuous road traffic noise is rightfully removed, mainly because of its low emergence and some differences in the emitted sounds. This explains why, after the ERF stage, urban environments are more favourable than rural contexts with road traffic: in rural areas, road traffic is not as dense and individual vehicles drive faster. The resulting sound resembles some train pass-by noise, explaining the difficulty to distinguish both sources.

The recall respectively decreases by 8, 12 and 5% in the 3 main contexts. Some expected train pass-bys are removed during the ERF stage, mostly because they do not meet the emergence criteria. Lowering the emergence threshold would retrieve these false negatives, but other false positives would also be included. Actually, these low-noise railway events do not matter much in the railway noise contribution estimation, which explains our threshold choice.

Finally, the F_1 is still driven by the precision. The ERF stage improves both metrics, but the area of improvement remains the false positives removal. As the most obvious ones have been removed during the ERF stage, another strategy is adopted to process the remaining events.

6. RAILWAY TRAINED CLASSIFIER (RTC)

Initially, the whole signal was classified frame-per-frame without any bias, and various environmental sound sources were expected. Now, the combined YTD and ERF steps provide vehicle-related emerging sound events. Building a classifier to apply on such events becomes easier, as the field of possible sound sources is smaller. This incites us to use a classifier trained on our own labelled events dataset. Training data would better resemble application data, which tends to improve accuracy. Our dataset contains 5 classes (Table 3): railway vehicles, roadway vehicles, construction machines, airborne noise and other surrounding noise. The first four classes are the most represented ones in the events provided after ERF. The last class gathers background noise and other rare events that may pass the YTD and ERF steps.

Table 3: Details of the training database used for the Railway Trained Classifier (RTC).

Category	Number of files	Total duration (s)
Railway vehicles	702	11290
Roadway vehicles	519	8559
Construction machines	896	8635
Airborne noise	720	7691
Other surrounding noise	1188	12481

In complex measurement environments, demarcation issues are sometimes met after the ERF step. Train pass-by demarcations are wider than the reference, and other events may also be affected in different ways. Therefore, the used classifier should be able to detect train pass-bys on portion of signals, instead of the whole event. This means that the event should be divided into frames prior to the classification. As in Figure 2, an event classification would result in a series of probability vectors, one per frame. Then, the series processing provides a decision to keep or reject the whole event.

The series processing consists in a frame-weighting process and a threshold application. Despite the demarcation issues, train pass-bys are expected to emerge from other sources noise on each event. Otherwise, they should not be included in the railway noise contribution estimation. We choose to weight each probability vector according to the normalized energy level of its corresponding frame. Mean value of the weighted probability vectors is computed and an event is categorized as train pass-by if its "Railway vehicles" score exceeds a determined empirical threshold.

The whole RTC process is defined, it remains to select the classifying neural network parameters. A balance has to be found for the frame duration: a longer frame contains more information but it should stay shorter than a complete train pass-by. The *YAMNet* frame length (960 ms) and overlap (50%) is fully adapted to our needs. No train pass-by lasts less than 960 ms and such extracts contain enough information for a neural network to identify vehicles (Section 4). Actually, as *YAMNet* parameters fully meet our requirements, we choose to perform transfer learning on the *YAMNet* network. The network convolution filters are frozen, and only the last dense layers will be updated to suit our database.

Going on with the same example as in Figures 3 and 4, the Figure 5 shows events that are classified as train pass-by or rejected. Each of the 6 events are well classified in this case. In particular, the freight train is correctly identified in spite of the wide demarcation. It occurred thanks to the energy-weighting processing that put forward frames during the pass-by against surrounding background noise.

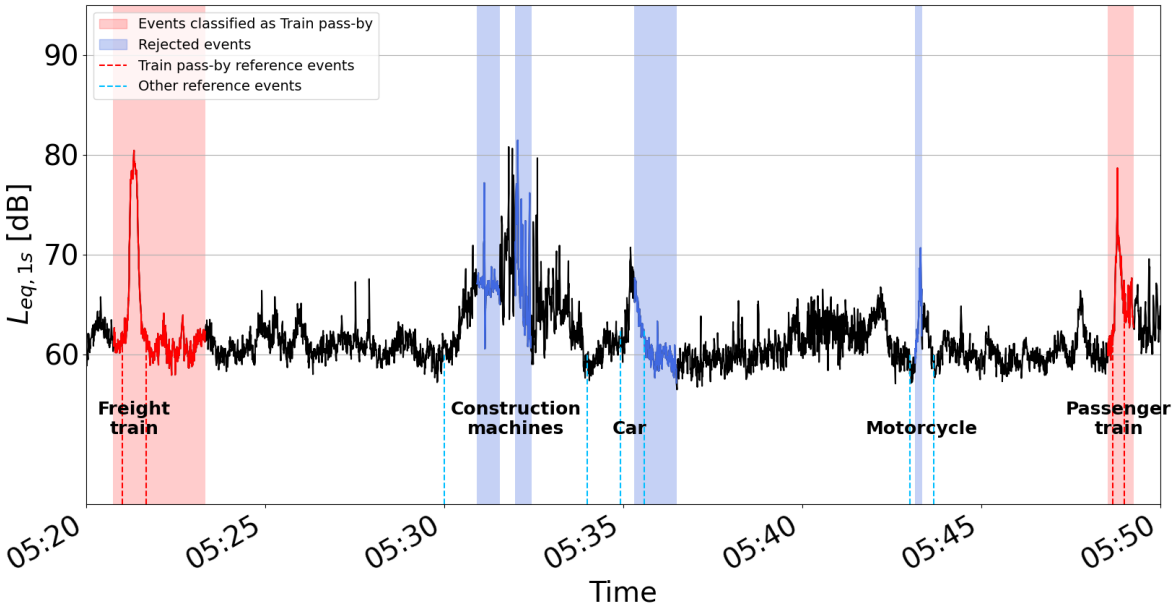

Figure 5: Evolution of equivalent sound level of a 30 minutes long audio measurement in urban context, with a representation of the ERF outcome.

The table 4 displays the RTC step outcome on the whole dataset. The RTC model accuracy is added. It computes the accuracy regarding the "Railway vehicles" class: $\frac{\mathrm{TP}_{rail}+\mathrm{TN}_{rail}}{\text{All predictions}}$, with TP_{rail} the correctly classified train pass-bys and TN_{rail} the correctly classified non-railway events.

Table 4: Evaluation of the RTC outcome on the whole dataset, separated in three different measurement contexts.

Context	**♯ of expected pass-bys**	**♯ of detected events**	**Prec. (%)**	**Recall (%)**	**F_1 (%)**	**RTC model accuracy (%)**
Rural w/o road traffic	249	228	92.1	84.3	88.0	96.1
Rural w/ road traffic	287	397	52.4	72.5	60.8	85.2
Urban	893	1767	45.8	90.7	60.9	73.5

In rural contexts, the high RTC model accuracy allows a significant precision increase, which even exceeds the recall. This means that there are more missed events than false positive, and the optimal performance in rural contexts would be obtained in lowering ERF and RTC thresholds. On the contrary, precision remains the lowest metric in contexts with road traffic. It is clear that the major difficulty regarding the task is the railway/roadway vehicles distinction. The greater risk of overlapping sources also explains the lower RTC model accuracy in urban context. Nevertheless, urban-made measurements are usually closer to the railway line, explaining the high recall. This means that raising ERF and RTC thresholds would enhance global accuracy in urban contexts.

Finally, the lower recall value in "Rural with road traffic" contexts is mainly due to a single measurement. Cicada song sound level is almost as high train pass-by noise during an extended period, leading to 25 missed events in a few hours. The recall would rise up to 81.2% without this measuring point. It shows that, even with encouraging results, the RTC model could still underperform in specific recording environment. Working on data augmentation and fine-tuning *YAMNet* convolution filters could be ways to improve the RTC model accuracy in such cases.

7. FINAL COMMENTS AND CONCLUSIONS

In this paper, a three-staged method is proposed to automatically detect train pass-bys on long-term raw audio measurements. From the outcome of the CNN *YAMNet*, frame-based vehicle-related sound events are first identified. Then, signal processing techniques and criteria are used to recompose events and start to separate train pass-bys from other sound events. Finally, to process remaining false positives, transfer learning with internal data is applied on the initial *YAMNet* model.

The succession of these 3 stages provides a train pass-by detection method whose precision depends on the measurement context. Our Railway Trained Classifier has a 73.5% accuracy in urban environments, which raises to 85% and 96% in rural contexts, respectively with and without road traffic. With little road traffic, more than 90% of detected events are indeed train pass-bys. However, precision stands around 50% on average whenever road vehicles interfere with the acoustic measurement.

Regarding the proportion of missed train pass-bys, it does not depend much on road traffic density but rather on the train model and measured sound level during the pass-by. *YAMNet* and our Railway Trained Classifier behaviours depend on the initial signal energy. Measurements in rural areas are usually further away from the railway line, and low noise train pass-by can be missed in the first detection (YVD stage).

Such missed events do not impact the railway noise contribution estimation. Railway noise contribution is estimated through several indicators, such as the railway $L_{day-evening-night}$ (average sound level of train pass-bys, with evening and night penalties). In 15 of the 17 different measurements, its estimation is less than 1 dB(A) away from the reference. Nonetheless, several areas of improvement have already been pointed out. A larger database could prevent issues in specific measurement contexts ; intelligent data augmentation techniques could also prevent some false negatives. Lastly, further work will focus on the classification models in order to propose classifiers fully adapted to the railway sound event detection.

REFERENCES

1. Xiuting Li, Qiu Dong, Boshen Wang, Haiyan Song, Shizhi Wang, and Baoli Zhu. The influence of occupational noise exposure on cardiovascular and hearing conditions among industrial workers. *Scientific Reports*, 9(1):11524, 2019.
2. Valéria Regecová and Eva Kellerová. Effects of urban noise pollution on blood pressure and heart rate in preschool children. *Journal of hypertension*, 13(4):405–412, 1995.
3. Stephen A Stansfeld and Mark P Matheson. Noise pollution: non-auditory effects on health.

British medical bulletin, 68(1):243–257, 2003.

4. Ministère de la Transition écologique et de la Cohésion des Territoires. Bruit, nuisances sonores et pollution sonore, 01 2021.
5. Nf s 31-088 : Acoustics - measurement of railway traffic noise in view of its characterization. Technical report, AFNOR, 1996.
6. Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2015.
7. Satvik Venkatesh, David Moffat, and Eduardo Reck Miranda. You only hear once: a yolo-like algorithm for audio segmentation and sound event detection. *Applied Sciences*, 12(7):3293, 2022.
8. Luis Antonio AZPICUETA-RUIZ, Miguel MOLINA-MORENO, Daniel DE LA PRIDA, and Antonio PEDRERO. An acoustic train pass-by detector based on parameters measured by sound analyzers: where machine learning meets human operator expertise. In *Proceedings of the 24th International Conference on Acoustics (ICA)*, pages 164–171, 2022.
9. Erwann Betton-Ployon, Abbes Kacem, Jerome Mars, and Nadine Martin. Analyse statistique de signaux acoustiques environnementaux pour la détection d'événements sonores. In *29° Colloque sur le traitement du signal et des images*, pages p. 1101–1104. GRETSI - Groupe de Recherche en Traitement du Signal et des Images, 2023.
10. Tony Marteau, Sitou Afanou, David Sodoyer, Sébastien Ambellouis, and Fouzia Boukour. Audio events detection in noisy embedded railway environments. In *European Dependable Computing Conference*, pages 20–32. Springer, 2020.
11. Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.