Sound source classification for soundscape analysis using fast third-octave bands data from an urban acoustic sensor network

Modan Tailleur, Pierre Aumond, Mathieu Lagrange, Vincent Tourre

This paper is part of a special issue on Advances in Soundscape: Emerging Trends and Challenges in Research and Practice.

Sound source classification for soundscape analysis using fast third-octave bands data from an urban acoustic sensor network

Modan Tailleur,[1, a] Pierre Aumond,[2] Mathieu Lagrange,[1] and Vincent Tourre[3]
[1] *Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France*
[2] *Univ Gustave Eiffel, CEREMA, UMRAE, F-44344 Bouguenais, France*
[3] *Nantes Université, École Centrale Nantes, CNRS, AAU, UMR 1563, F-44000 Nantes, France*

(Dated: 23 August 2024)

The exploration of the soundscape relies strongly on the characterization of the sound sources in the sound environment. Novel sound source classifiers called Pre-trained Audio Neural Networks (PANNs), are capable of predicting the presence of over 500 diverse sound sources. Nevertheless, PANNs models use fine Mel spectro-temporal representations as input, whereas sensors of an urban noise monitoring network often record fast third-octaves data which have significantly lower spectro-temporal resolution. In a previous study, we developed a transcoder to transform fast third-octaves into the fine Mel spectro-temporal representation used as input of PANNs. In this paper, we demonstrate that employing PANNs with fast third-octaves data, processed through this transcoder, does not strongly degrade the classifier's performance in predicting the perceived time of presence of sound sources. Through a qualitative analysis of a large-scale fast third-octave dataset, we also illustrate the potential of this tool in opening new perspectives and applications for monitoring the soundscapes of cities.

[https://doi.org/10.1121/10.0026479]

[Editor: Francesco Aletta] Pages: 1–12

I. INTRODUCTION

Over the years, noise level measurements and simulations have been widely used to gauge noise nuisance caused by different urban environments. Directive 2002/49/EC of the European Union (Commission, 2022) enforces noise level mappings in urban areas with more than 100,000 residents and close to major transportation hubs. However, the assessment of noise extends beyond mere quantitative loudness measures, as it is inseparable from human perception (Lavandier and Defréville, 2006). The acoustic environment as perceived by humans in a specific context, is encapsulated by the definition of "soundscape", as standardized in ISO 12913-1:2014 (ISO, 2014). Gaining insight into human emotions, such as annoyance experienced in a sound environment, proves to be non trivial and dependent on the specific subgroups of citizens involved (Tarlao *et al.*, 2021; Yong Jeon *et al.*, 2011). For instance, the sound environment near a bar perceived by residents may not be as pleasant as it is by tourists, due to their distinct expectations. These conflicting assessments highlight the need for more nuanced approaches in soundscape evaluation that go beyond a singular reliance on evaluating annoyance.

Analyzing the various sound sources within an environment provides a holistic understanding of soundscape quality. In previous studies, road traffic, human voices, and bird sounds have been frequently considered in soundscape evaluations (Aumond *et al.*, 2017; Ricciardi *et al.*, 2015), representing the prevalent sounds in mechanical, human, and natural urban environments (ISO, 2014; Jeon and Hong, 2015). Previous work argues that as opposed to sole reliance on loudness assessment, focusing on characterizing these three environments is essential to better describe a soundscape (Aletta *et al.*, 2016; Axelsson *et al.*, 2010; Botteldooren *et al.*, 2011).

Sound sources are usually evaluated by their presence, sound level, or dominance in the sound environment. In the soundscape standard ISO/DIS 12913-2 (ISO, 2018), the term "dominance" of a source is employed, implying a notion of competition among sources. In this paper, we align with the terminology employed by Lavandier et al. (Aumond *et al.*, 2017; Gontier *et al.*, 2019; Lavandier *et al.*, 2021), and adopt the term "time of presence" instead. The significance of this choice will be elaborated further in Section V.

[a] modan.tailleur@ls2n.fr

To evaluate the time of presence of sound sources, researchers frequently employ questionnaires, either *in situ*, or *in vitro*, *i.e.* by having participants listen to sound recordings using headphones. These questionnaires are typically conducted on small audio segments of durations ranging from 30 seconds to several minutes (Aumond *et al.*, 2017; Axelsson *et al.*, 2010; ISO, 2014; Papadakis *et al.*, 2023). This approach yields valuable insights on the perception of soundscapes. However, human annotations are time-consuming and error-prone, therefore infeasible to scale to large datasets.

Acoustic measurements present a promising solution to address this challenge, provided they can effectively help predict the presence of sound sources. The IEC 61672-1 (Commission *et al.*, 2013) standardizes the measurement of fast (125-ms hops) and slow (1-s hops) third-octave spectral representations, which finds application in various noise monitoring contexts (Aumond *et al.*, 2017; Can *et al.*, 2021; Farrés, 2015; Mietlicki *et al.*, 2015; Nilsson *et al.*, 2007; Torija *et al.*, 2013). The characteristics of the fast third-octave measurements provided by Cense sensors (Ardouin *et al.*, 2021) are described in Table I. Fast third-octave spectrograms present several advantages in long-term monitoring applications. In particular, they enforce unintelligibility, thus preserving privacy, as demonstrated by Gontier et al.(Gontier *et al.*, 2017). Furthermore, these representations are lightweight, with a bit rate 140 times lower than that of mono waveform recordings (16-bit - 32 kHz), and 30 times lower than that of 64 Mel bands with 10ms hops. They also facilitate the deployment of affordable noise monitoring networks, as most sound level meters already provide these acoustic features.

Some acoustic indices, which can be computed from third-octave measurements, have shown significant correlations with the presence of specific sound sources, such as LAeq for road traffic and the Time Frequency Second Derivation Index (TFSD) for bird presence (Aumond *et al.*, 2017). The package seewave (Sueur *et al.*, 2016) incorporates many other acoustic indices, designed for ecoacoustic analyses, which are intended to detect geophonic, anthropogenic, and natural sounds. Unfortunately, creating specific acoustic indices is time-consuming and case-specific as it requires a context-aware understanding of the acoustic characteristics of the target sound source. For example, while TFSD effectively captures rapid spectro-temporal variations in bird sounds within a particular frequency range, it does not perform optimally when applied to other sound sources with similar characteristics yet different frequency ranges, such as human voices (Aumond *et al.*, 2017). Similarly, although LAeq correlates strongly with the time of presence of traffic noise in busy streets, its effectiveness in predicting the presence of road traffic diminishes in environments dominated by other sound sources, such as construction sites or factories.

Artificial intelligence (AI) has proven to be a valuable tool for detecting sound sources (Bansal and Garg, 2022) and predicting their perceived time of presence. Gontier at al. (Gontier *et al.*, 2019) obtained accurate predictions of the perceived time of presence of traffic, voices and birds, by training a Convolutional Neural Network (CNN) on fast third-octave spectrograms. In this paper, we refer to their model as **CNN-TrainSynth**. While CNN-TrainSynth shows good performance on the Cense Lorient dataset (Can *et al.*, 2021), it also demonstrates a lack of robustness when applied to alternative datasets recorded with fast third-octave. This limitation stems, in part, from the model's training on highly homogeneous data. Furthermore, despite the standardization of time weighting for fast third-octave spectrograms, variations in other parameters, such as the number of frequency bins and the frequency bounds, can induce problems in generalization to other data distribution. Additionally, while the architecture could easily be adapted to a given sound source, it would require re-training and thus a cumbersome procedure to produce a clean and annotated dataset.

As mentioned previously, annoyance assessment is inherently intertwined with the listening context. Predicting annoyance level with a trained deep learning model would thus pose serious problems in interpretation and transparency for users, a practice that differs from the principles of explainable AI (Mueller *et al.*, 2019). In our case, we have chosen to focus on time of presence, which is a perceptual concept that is less sensitive to context. Addressing the same concerns, Hou et al. (Hou *et al.*, 2023) aimed to develop a model that jointly predicts sound sources and annoyance, exemplifying the ongoing pursuit of enhanced explainability and precision in predicting the emotional impact of the sound environment.

spectral representation	10-ms 64Mel	fast third-octave
origin	PANN	Lorient Cense Network
sample rate	32kHz	32kHz
window size	1024 (32ms)	4096 (128ms)
fft size	1024 (32ms)	4096 (128ms)
hop size	320 (10ms)	4000 (125ms)
window	hann	tukey
frequency bins	64	29
min frequency	50Hz	20Hz
max frequency	14kHz	12,5kHz
mel normalisation	slaney	-
mel formula	slaney	-
bit rate	100kb/s	3,71kb/s

TABLE I. Differences between PANN (ResNet38) and Cense spectral representations

In recent years the realm of deep learning has seen a surge of a family of robust and versatile pre-trained classifiers, known as Pre-trained Audio Neural Networks (PANNs) (Kong *et al.*, 2020). PANNs are highly effective deep learning models that have been trained on Au-

	acoustic indicator	CNN-TrainSynth	PANN-1/3oct	PANN-Mel
input spectrogram	fast third octave	fast third octave	fast third octave	10-ms 64Mel
output classes	1	3	527	527
needs training annotations	Yes	Yes	No	No

TABLE II. Specifications of the different inference methods. The need for training annotations refers to the need to use time of presence annotations for each sound source

dioset, an extensive audio dataset comprising more than 2 million audio clips (Gemmeke *et al.*, 2017). These models operate as sound classifiers and are capable of making predictions on the presence of 527 different sound sources without the need for additional training. For each class, they output a confidence score between 0 and 1, indicating the likelihood of the source's presence. Over 20 distinct pre-trained models are accessible online, all employing the same 10-ms 64Mel bands spectro-temporal representation as input. This representation has a much higher frequency and time resolution compared to fast third-octaves, as shown in Table I.

This paper builds upon the work presented by Tailleur et al. (Tailleur *et al.*, 2023) which focused on introducing an algorithm for transforming fast third-octave representations into 10-ms 64Mel bands representations. This transcoding technique allows the use of classification algorithms that necessitates Mel spectro-temporal representations as input, such as PANNs, with fast third-octave band measurements. In the following sections, the ResNet38 PANN model with 10-ms 64Mel bands spectrograms as input will be referred to as **PANN-Mel**, and the same PANN model that uses transcoded Mel spectrograms from fast third-octave measurements will be referred to as **PANN-1/3oct**. A summary of the available models and methods to predict the time of presence of sound sources are presented in table II. Technical aspects of the transcoder are summarized in section II. Despite only being tested for this specific use case scenario, theoretically the transcoder can be used to convert any spectro-temporal representation into 10-ms 64Mel bands spectrograms. It was demonstrated that the transcoding technique applied to PANN on fast third-octaves enabled satisfactory classification performances on two urban sound datasets: SONYC-UST and UrbanSound8k. However, the assertion that PANN-1/3oct output classes are highly correlated with perceptual evaluations of time of presence is yet unproven.

This study is specifically focused on showcasing the effectiveness of PANN in accurately predicting the perceived time of presence, a crucial metric in soundscape analysis, for diverse sound sources. It particularly assesses the applicability of PANN in analyzing the time of presence of datasets recorded with fast third-octave representations. PANN predicts accurately not only the time of presence for traffic, voices, and birds but also the presence of a diverse range of 527 different sound sources. Thanks to the transcoder developped by Tailleur et al. in (Tailleur *et al.*, 2023), using PANN on fast third-octave data allows predictions on lightweight datasets, while ensuring privacy. Consequently, this approach facilitates analysis on large datasets, such as Lorient Cense (explored in section V) which contains more than 500k hours of fast third-octave spectro-temporal data on 75 different sensors.

In the forthcoming section II, we present the transcoding algorithm used for PANN-1/3oct. In section III, we elaborate on our findings concerning perceived time of presence assessment performance. We then show an analysis of the soundscape of Lorient through a fast third-octave recorded database in section IV. In section V, we will discuss the relevance, the opportunities, and potential enhancements arising from this approach. Open source code is available at `https://github.com/modantailleur/paperSoundscapeSourceClassification`.

II. TRANSCODER

PANNs models (Kong *et al.*, 2020) take a 10-ms 64Mel frequency band spectrogram as input, as shown in table I. Audio can easily be transformed into the corresponding Mel spectrogram in order to use PANN. Nonetheless, there is no easy transformation from fast third-octave spectrograms to 10-ms 64Mel frequency band spectrograms, as the latter have much finer resolutions. In order to use PANN models to predict the presence of sound source from fast third-octave recordings, a transcoder is employed to convert them into 10-ms 64Mel frequency bands spectral representations. In the subsequent sections, we will refer to the fast third-octave spectrograms as **"coarse spectrograms"** and the 10-ms 64Mel frequency bands spectrograms as **"fine spectrograms"**. We direct readers to (Tailleur *et al.*, 2023) for further details on the training procedure, and on the model's classification performances compared to other state-of-the-art methods.

A. General description

The transcoder outputs the fine spectrogram format that is required as input of PANNs models. This transformation is performed on audio segments with a duration of 1 second. The Lorient Cense project's method for fast third-octave (coarse spectrogram) calculation (Can *et al.*, 2021) is chosen. This method involves computing 29 third-octave bands within the frequency range of 20Hz to 12.5kHz, using a tukey 125-ms temporal window, as depicted in table I.

The transcoder is trained on the TAU Urban Acoustic Scenes 2020 Mobile dataset (Mesaros *et al.*, 2018), an urban dataset of raw audio data. We highlight that this dataset does not include any annotations regarding

the presence of sound sources in the audio recordings. It consists of 30h of 10-second audio clips from 10 different acoustic scenes.

B. Transcoder architecture

The proposed CNN transcoder model consists of two parts: a Pseudo-INVerse (PINV) transcoder and a Convolutional Neural Network (CNN), as shown in Figures 1 and 2. The PINV transcoder presented in Figure 1 first reconstructs the full-band spectrogram from the coarse spectrogram using a pseudo-inverse method (Penrose, 1955). The PINV stage enables leveraging knowledge from the third-octave transform to provide an initial estimation of the fine spectrogram. This stage also theoretically allows the model to adapt to any coarse spectrogram calculated using different numbers of frequency bins or varying time weightings. The CNN stage improves the quality of this estimation by adding residual information to it, as illustrated in Figure 2. The model is very light, representing 0.3% of our selected ResNet38 PANN's total number of parameters.

C. Learning approach

A teacher-student approach consists in leveraging an existing pre-trained model to distill its knowledge into a student model that learns from the teacher's output (Hinton *et al.*, 2015). This approach offers several advantages over traditional supervised learning, where a model is trained solely from human-annotated audio samples. By exclusively relying on the outputs of the teacher model, a teacher-student training eliminates the need for human annotations. Consequently, it is adaptable to various audio datasets, making it highly convenient for training the model with a large quantity of audio samples. As a result, such a training method often produces more robust models.

The teacher-student approach taken to train the transcoder diverges slightly from the aforementioned regular teacher-student methods, while benefiting from the same advantages. Notably, only the parameters of the CNN transcoder are updated during training (see Figure 3), whereas the entire student model would be trained if a regular teacher-student approach had been employed. As a result, at the end of the transcoder training, the PANN model used in the student process remains identical to the one employed by the teacher. Training of the transcoder is performed using the Binary Cross-Entropy (BCE) loss function, computed on the predictions of both the teacher and student PANN classifiers.

Several extra benefits arise from this improved teacher-student approach. First, it minimizes computational complexity by limiting training to only a fragment of the student network. Additionally, this approach improves versatility, allowing the model to be easily adapted to a broad spectrum of pre-trained classifiers that utilize similar fine spectrogram as inputs.

To provide intuitive insight on this innovative teacher-student process, the transcoder learns to reconstruct a spectrogram based on the high-level features resulting from the 527 output classes of PANN-Mel. Instead of aiming for a perfect reconstruction of the target, the model thus prioritizes crafting a spectrogram that encapsulates the fundamental characteristics of a sound source being present or absent. In fact, the transcoder's objective is not an exact retrieval of what was lost but rather a reconstruction of coherent data yielding predictions close to those of the PANN-Mel teacher model.

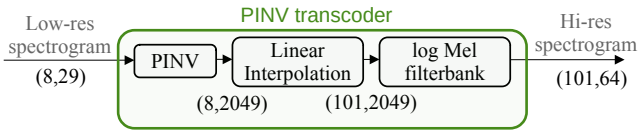

FIG. 1. PINV transcoder architecture, to recover a 1s sample fine spectrogram from a 1s sample coarse spectrogram

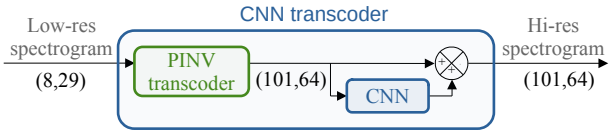

FIG. 2. CNN transcoder architecture, to recover a 1s sample PANN fine spectrogram from a 1s sample coarse spectrogram

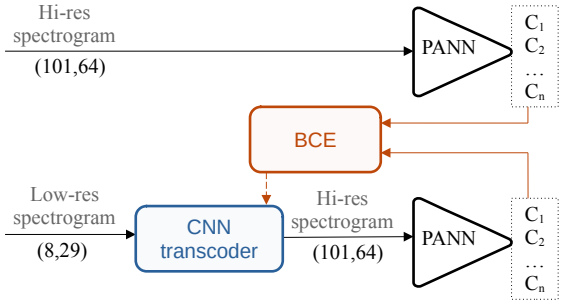

FIG. 3. PANN CNN transcoder trained with a teacher-student approach, using the Binary Cross-Entropy (BCE) loss function

D. Validation

The training of the model lasts approximately 4h on a V100 GPU. Interestingly, the resulting PANN-1/3oct model trained with a transcoder, outperforms those that are trained with a regular teacher-student method. This surprising outcome is particularly noteworthy considering that the CNN transcoder is substantially smaller in

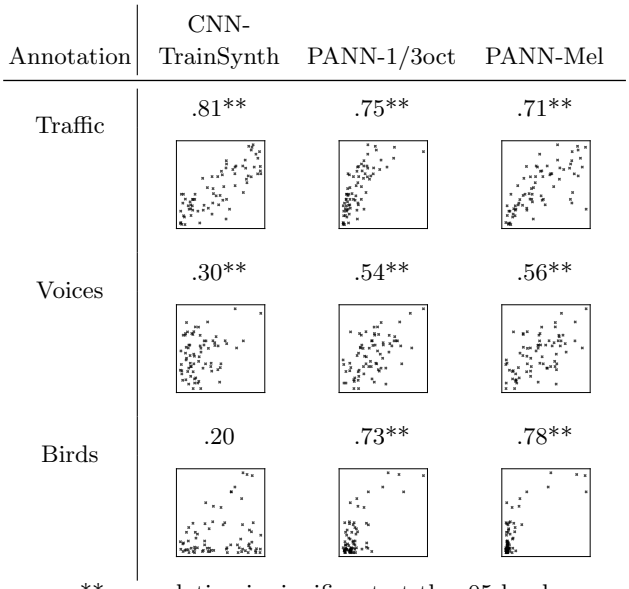

FIG. 4. Pearson correlation on GRAFIC. Each plot shows the annotations (y-axis) depending on the predictions (x-axis) (n=74).

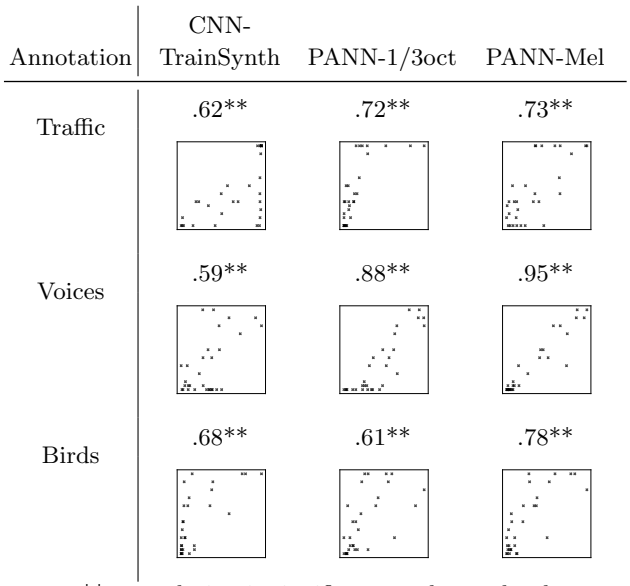

FIG. 5. Pearson correlation on Lorient-1k. Each plot shows the annotations (y-axis) depending on the predictions (x-axis) (n=30).

size. PANN-1/3oct obtains an accuracy of 89.3% in predicting the same first class as the teacher PANN-Mel model on the evaluation dataset of TAU Urban Acoustic Scenes 2020 Mobile dataset (Mesaros *et al.*, 2018). Models trained using regular teacher-student methods only reached a maximum accuracy of 83.7%. The proposed model has also been tested in multi-class classification tasks and multi-label classification tasks, reaching a 62.4% accuracy on the UrbanSound8k dataset, and a .44 mAUPRC on Sonyc-UST dataset, see (Tailleur *et al.*, 2023) for more details. Even if this model has proven good performances for classification tasks, it's performances for assessing the time of presence of different sound sources remains to be evaluated.

III. QUANTATIVE PERFORMANCE ANALYSIS

A. Materials and Method

1. Datasets

Two datasets of audio recordings are used for evaluating the ability of PANN-1/3oct in predicting the perceived time of presence of traffic, voices and birds: the Lorient-1k and GRAFIC datasets.

Lorient-1k (Gontier *et al.*, 2021) is recorded in the city of Lorient in France with a zoom H4N at 10 different locations. It has a total duration of 22.5min, consisting in 30 acoustic scenes of 45s each. Four experts, who are members of the project team with significant expertise in soundscape analysis, annotated their perceived time of presence of birds, traffic and voices over the 45s. Due to the low number of annotators, the most experienced expert homogenized the dataset.

The GRAFIC dataset (Aumond *et al.*, 2017) was created from a soundwalk in the 13th district of Paris. The soundwalk was designed to cover a wide variety of urban sound environments. The recording system employed was an ASAsense device, which was mounted on the operator's backpack as researchers were walking alongside the volunteering citizens. The annotation process involved a total of 37 different participants, with 9 to 11 of them present simultaneously at each 19 location. While traversing different locations, they completed questionnaires detailing the pleasantness, liveliness, perceived loudness of specific sound sources (such as mopeds, cars, horns, trucks, and buses), and the perceived time of presence of others (including traffic, voices, footsteps, birds, wind, and water). With four sessions conducted, 74 audio files were obtained from the 19 different locations, ranging in duration from 1 to 3 minutes each. An estimation of the perceived time of presence for the various sound sources was obtained by averaging the questionnaire results from annotators present during each session at each location.

2. PANN prediction processing to predict the perceived time of presence

PANN ResNet 38 model processes Mel spectrograms segments with a duration of 10-s, and outputs a confidence level between 0 and 1 for the presence of each 527 different sound sources. To derive an indicator for the presence of a specific source, we computed the average of the model's outputs across all available 10-s chunks for each output class.

To represent the time of presence of traffic, voices, and birds, we select three output classes from PANN that closely align semantically with these sources, namely "traffic noise, roadway noise", "speech", and "bird vocalization". For further details on the available PANNs output classes, please refer to the Audioset ontology (Gemmeke *et al.*, 2017).

PANN's original training dataset is based on normalized audio. However, normalizing each audio file individually could potentially lead to misleading results, especially for relatively silent audio files that lack significant level dynamics. In response, we adopt a strategy of normalizing the audio data to the maximum level across the entire dataset, with an exception for the top 1% percentile of the highest levels. This normalization process thus simply involves adding the same level offset to all audios. The exclusion of the highest levels mitigates the impact of outliers (e.g., sensor hit, wind in the microphone), contributing to improved model performance.

On a GPU V100, PANN-1/3oct processes 175 seconds of real-time audio within a single second.

B. Results

The correlation results for the three selected PANN-1/3oct classes are presented in Figures 4 and 5, alongside the correlations for the CNN-TrainSynth and the PANN-Mel models. PANN-1/3oct demonstrates overall superior performance compared to the CNN-TrainSynth model for the 3 sound sources. It only exhibits slightly lower performance compared to the CNN-TrainSynth model on traffic for GRAFIC, and got similar correlations for birds on Lorient-1k. Furthermore, the performance of PANN-1/3oct closely rivals that of PANN-Mel, despite PANN-Mel utilizing more refined spectro-temporal representations as input. PANN-1/3oct appears to outperform PANN-Mel for traffic correlation on Lorient-1k. However, this observation is likely influenced by the low number of data points on which the correlation is calculated.

IV. QUALITATIVE PERFORMANCE ANALYSIS

Previously exploited datasets such as Lorient-1k and GRAFIC are relatively small, encompassing less than 4h of audio recordings in total. Unfortunately, no larger-scale audio datasets in the literature include annotations for the time of presence of sound sources. Due to this limitation, we now focus on a large-scale dataset of fast third-octave recordings that doesn't contain any explicit annotations. We conduct a qualitative analysis of this dataset to assess if the PANN-1/3oct prediction results align with our expectations.

A. Methods

1. CENSE sensor network

The CENSE project has developed low-cost noise monitoring sensors designed to be incorporated into a large network of sensors (Ardouin *et al.*, 2021; Picaut *et al.*, 2020). Such a network of sensors has been deployed between January 2020 and April 2022 in the city center of Lorient in France. The overall network includes 78 noise sensors connected to the cloud. This network utilizes the public street lamp network, incorporating power-line communication systems.

The sensors utilize micro-electromechanical system (MEMS) microphones, equipped with a Raspberry Pi for recording and transmitting purposes, featuring real-time audio processing capabilities. They allow the recording of an acoustic spectrum every 125 milliseconds using 29 third-octave bands covering the frequency range from 20 Hz to 12.5 kHz (as shown in table I).

2. Dataset processing

The entire dataset encompasses over 500k hours of fast third-octave measurements derived from the previously outlined network. Given the computational efficiency of PANN-1/3oct, estimating source levels across the entire dataset would necessitate about 3k hours of computation. Consequently, we opt to sample the dataset within specific time intervals and spatial areas to showcase the efficacy of PANN-1/3oct in the context of large-scale urban datasets.

For long-term analysis on multiple sensors, we randomly select a certain number of 1-min samples per day of calculation. In cases where data is not accessible on a particular day or for a randomly selected sensor, that specific sample is omitted from the subset. For short-term analysis on only a few sensors, the entirety of the available 1-min samples per day are selected. The time period of each study (TP), the number of samples used per day (NSD), the total number of samples (TNS), and the number of sensors available or involved (NS) are specified within the figures' captions of each analysis.

As detailed in section III A 2, the subsets are normalized using the maximum level observed within the dataset of 33,443 recordings collected between January and March 2020, with the top 1% percentile excluded. This established level offset is consistently applied to all other subsets of the Cense Lorient dataset.

3. Predictions processing

The PANN model predictions are not standardized. For instance, the predictions for the class "Bird Vocalization" in GRAFIC and Lorient-1k, discussed in section III, fall within the range of 0 to 0.2, while the corresponding annotations for each traffic, voice and birds annotations span from 0 to 1. This annotation scale of 0 to 1 represents a linear and bipolar scale, ranging from "sound source not present at all" to "sound source always present". Similar discrepancies are observed in traffic predictions (0 to 0.4) and voice predictions (0 to 0.6). To address this disparity and ensure alignment with perceived time of presence annotations, we employ linear interpolation without intersection, calculated on the GRAFIC and Lorient-1k predictions. This approach ef-

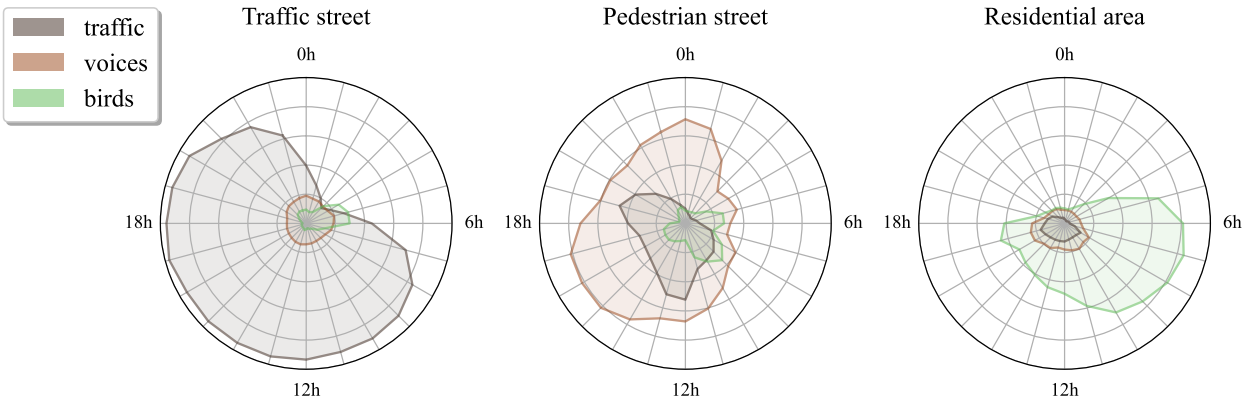

FIG. 6. Clock graphs representing the estimated perceived time of presence of traffic, voices and birds, on a sensor close to a traffic street, a pedestrian street and a residential area. The source time of presence prediction is averaged per hour. The time period considered ranges from January 1st to March 1st of 2020, with a random data sampling rate of 14% per day across 67 sensors, resulting in a total of about 500 hours of recording.

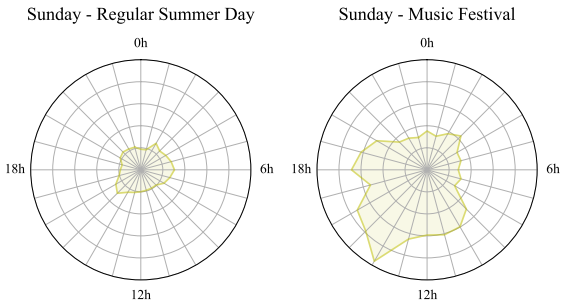

FIG. 7. Clock graphs representing the presence of music during a music festival, and on a regular summer day. The source presence likelihood is averaged per hour. The time period considered for Figure a) encompasses every sunday of July 2021, taking into account all data from 5 sensors, resulting in a total of about 200 hours of recording. The time period considered for Figure b) is the 8th of August 2021, taking into account all data from 5 sensors, resulting in a total of about 100 hours of recording.

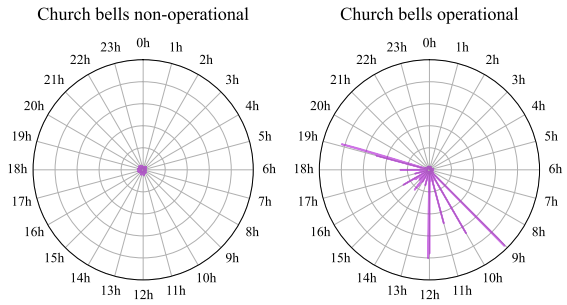

FIG. 8. Clock graphs representing the presence of church bells on January 2020 when the bells were operational, and on October 2020, when they were not. The source presence likelihood is averaged per minute. The time period considered for Figure a) is October 2020, taking into account all data from 1 sensor, resulting in a total of about 300 hours of recording. The time period considered for Figure b) is January 2020, taking into account all data from 1 sensor, resulting in a total of about 600 hours of recording.

fectively adjusts traffic, voices, and birds predictions to a range consistent with perceived time of presence values, enabling fair inter-class predictions comparisons.

Regrettably, this normalization procedure cannot be extended to other classes like "music", "church bell", and "civil defense siren" due to the unavailability of datasets with annotations regarding the time of presence for these sound sources. Instead, the PANN-1/3oct output for those classes will systematically be compared to another time period or another location.

B. Results

1. Temporal analysis

In order to evaluate the accuracy of the model in predicting the time of presence of traffic, voices and birds, we consider the time period between January 1, 2020, and March 1, 2020. To delve into the temporal analysis of this time period, we focus on three specific sensors, each selected to represent distinct acoustic environments. One sensor is in a residential area near a children's playground, one on a busy traffic street, and one situated in a pedestrian area with bars and nightclubs. The intention behind this selection is to encompass various temporal

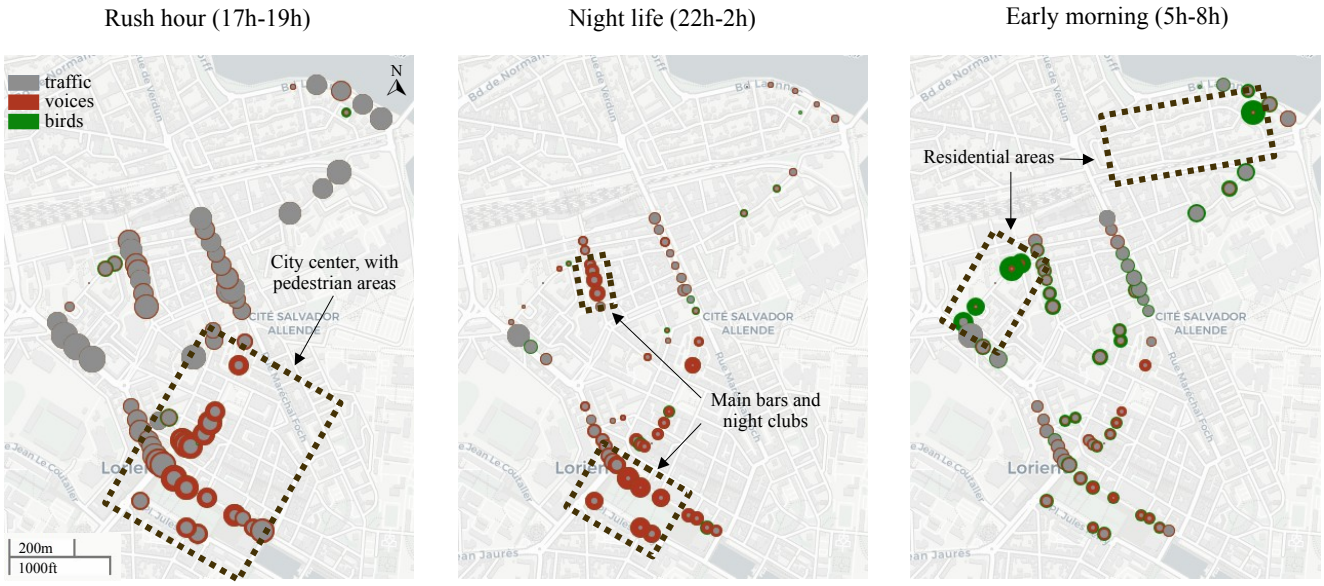

FIG. 9. Maps representing the time of presence of traffic, voices and birds over different time intervals. Each node has 3 colors, the area of each color being related to the time of presence of their respective sound source. The circles are not superposed, but should be seen as one circle with the area of each color within it representing the time of presence of each sound source. The time period considered ranges from January 1st to March 1st of 2020, with a random data sampling rate of 0.7% per day across 67 sensors, resulting in a total of about 600 hours of recording.

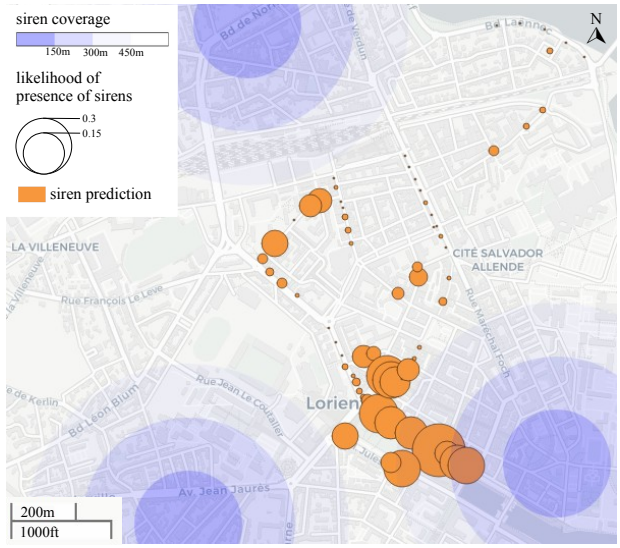

FIG. 10. Map illustrating the predicted presence of civil defense sirens on the first Wednesday of each month at 11h45. The map displays the locations of sirens and their estimated coverage. The time period considered encompasses every first wednesday of the month at 11h45 between January and July of 2020, taking into account all data from 67 sensors, resulting in a total of about 5 hours of recording.

patterns of the three sound sources across these three distinct sensor locations.

We observed highly contrasting distributions of traffic, voices, and birds (see Figure 6), all in line with our expectations:

- In the traffic street, a notably higher presence of traffic is observed compared to the other locations, particularly from 7h to 19h. Additionally, we identify the presence of birds on the traffic street, with their activity commencing approximately one hour before sunrise, as anticipated. The existence of trees along this street provides a plausible explanation for the birds' presence. The decrease in bird activity during the rest of the day could be attributed to birds leaving, reduced chirping, or being masked by traffic noise.

- In the pedestrian street, voices become prevalent in the afternoon, specifically starting from 14h. This voice activity declines from 18h to 21h and increases between 21h and 0h. This pattern aligns with the proximity of this sensor to bars that usually open in the afternoon and close around 2h. The surprising fluctuations in the time of presence predictions for birds throughout the day on the pedestrian street could be attributed to the presence of various other sound sources in this area (e.g. speech, bicycle bell, horn), leading to inaccuracies in the predictions.

- Within the residential area, birds are consistently more active throughout the day, with a notable concentration in the morning, two hours before sunrise. The predictions also show a slight increase at sunset, aligning with expectations. Additionally, we observe minor fluctuations in voice activity between 7h and 8h, and between 17h and 18h, likely linked to people leaving and returning home.

PANN-1/3oct also demonstrates effectiveness in predicting the presence of other sound sources. Figure 7 illustrates a substantial increase in music presence associated with the "music" PANN-1/3oct class during a festival day compared to a regular sunday. The "Festival Interceltique de Lorient", an outdoor music event, exhibited a substantial concentration of music on the Sunday 8th of July 2021, primarily in the afternoon. Of particular note is the prominent main outdoor concert that commenced at 14h30, during which musicians paraded through the streets. The beginning of this parade correlates with the heightened music activity predicted by PANN-1/3oct in this time period.

Furthermore, Figure 8 illustrates the temporal patterns associated with the "church bell" PANN class. During January 2020, when the bells were in full operation, PANN-1/3oct predictions show that they rang in the opening minutes of each hour. In October 2020, the bells of Notre-Dame-De-Victoire were not ringing due to the structural damages caused by the vibrations produced by the bells. PANN-1/3 oct has indeed extremely low predictions for the church bell class on this time period. The predictions for church bells are significantly higher at 9h, 12h05, and 19h05. This is probably due to the Angelus, a daily Catholic prayer traditionally recited three times a day, usually at 7h05, 12h05, and 19h05. In certain churches, the 7h05 prayer may be postponed to a later hour, which seems to be 9h for the Notre-Dame-De-Victoire church.

2. Spatial analysis

Throughout the entire sensor network, we identify distinct patterns in sound behaviors based on the general sensor location and time periods between January 1, 2020, and March 1, 2020 (see Figure 9). Notably, boulevards exhibit significantly higher levels of traffic sounds compared to other street types. Pedestrian streets and areas near shops and bars tend to have a greater prevalence of voices. During nightlife periods, the sensors close to regular shops tend to feature lower levels of voices than the sensors close to bars and nightclubs. Residential areas show an increased presence of birds, coupled with a low presence of voices and traffic, particularly in the morning, aligning with the expected tranquility of nearby parks. Furthermore, our analysis reveals that voice activity predominantly occurs during rush hours and nightlife periods, being notably absent during the early morning. In contrast, traffic sounds are consistently present across all sensors, except during the night hours when the city experiences an overall decrease in activity.

In France, civil defense sirens are activated every first Wednesday of the month precisely at 11h45 These sirens are distributed across various locations in the city. Three main sources are situated in the city center of Lorient. Figure 10 illustrates the prediction of "Civil Defense Siren" presence for each first Wednesday of the month between January 2020 and July 2020, at 11h45. In comparison to this specific time, the predictions for the presence of sirens for the remainder of the day (when civil defense sirens are not activated) are negligible, being 70 times lower on average. During siren activation, the prediction scores tend to be higher for sensors located closer to the siren sources.

V. DISCUSSION

A. Why use third-octaves measurement ?

While it's possible to argue for an alternative approach involving sensors that capture raw audio, make predictions on sound source presence at regular intervals, and send only the predictions to a server to address privacy concerns, there are compelling reasons to consider predicting sound sources directly from third-octave measurements. Storing predictions rather than third-octaves necessitates sensors capable of making such predictions. This implies a requirement for more advanced, potentially higher-cost sensors. Moreover, relying solely on predictions would mean that future access to the original audio data would be lost. In this context, predicting from third-octave data with ever improving AI models could potentially lead to more efficient predictions in the future, even considering that we have access to raw audio in this day and age. Furthermore, as third-octave measurements from sound level meters are normed (Commission *et al.*, 2013), there are many already recorded datasets where PANN-1/3oct could be used to make sources predictions (Farrés, 2015; Mietlicki *et al.*, 2015; Nilsson *et al.*, 2007; Torija *et al.*, 2013).

B. Why use perceived time of presence? Isn't evaluating the time of presence the same as evaluating the signal duration of a sound source ?

The evaluation of the presence of sound sources in human assessments is often approached using the concept of "dominance," which aims to capture the perceived level of each sound source (Axelsson *et al.*, 2010; Hong and Jeon, 2015; Mitchell *et al.*, 2021). We believe that the term "dominance" can be meaningful when multiple sound sources are present in a scene. Most studies suggest measuring dominance as a score between 0 and 10. However, in our view, this approach has certain limitations. Implicitly, asking for the "dominance" of a source implies that other sources are being dominated. Therefore, evaluating dominance should involve more of a comparative assessment (e.g., sound source A is twice as dominant as sound source B, which is equally dominant as sound source C), rather than providing individual

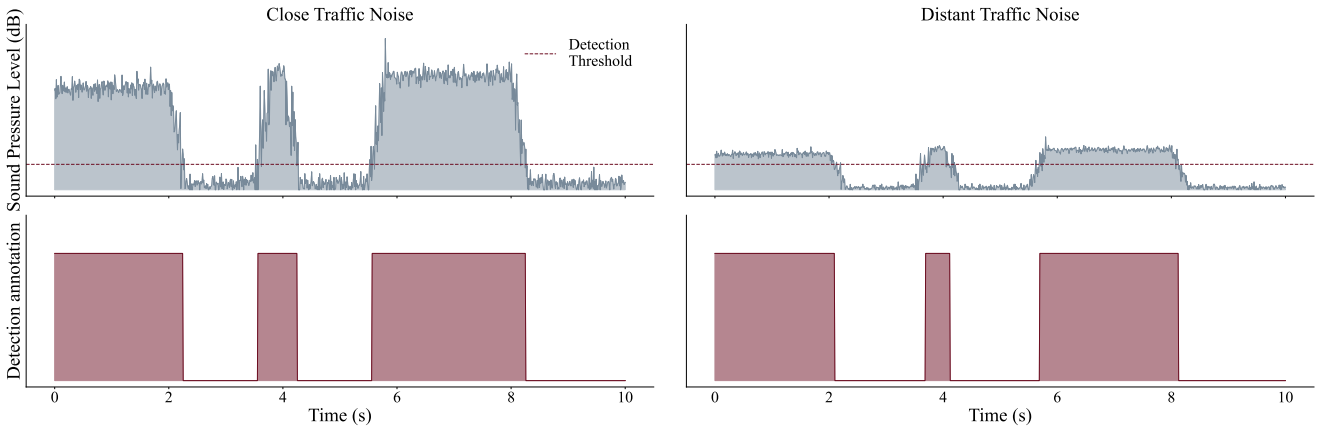

FIG. 11. Differences between time of presence annotation procedure, and mean of detection annotation on a synthetic example. On this example, close traffic noise would have a perceived time of presence close to 1/1, and distant traffic noise a perceived time of presence close to 0/1, even if their mean of detection annotation would be extremely close.

scores for each sound source (e.g., each sound source has a dominance score between 0 and 10). Consequently, the term "dominance" has been replaced with "time of presence" in some soundscape evaluation studies (Aumond *et al.*, 2017; Gontier *et al.*, 2019; Lavandier *et al.*, 2021), a choice also made in this article.

It is crucial to acknowledge that the concept of "time of presence" comes with inherent limitations and should not be confused with the precise determination of the duration of a sound source being active in a given environment, which would require specifying exact onsets and offsets. The term "time of presence" functions as a straightforward lexical substitution for the perceived overall level of a sound source (Lavandier and Defréville, 2006), suggesting a holistic assessment of its presence within a soundscape. Notably, a sound source may receive a maximum time of presence score while being physically present above a certain audible threshold only half of the time, as human perception would naturally fill gaps in auditory information during a holistic evaluation. Similarly, a source physically present and audible for a significant duration, but at a low level or very large distance, would result in a perceived time of presence score close to none. Consequently, datasets annotated with onsets and offsets of sound sources, such as Singa::Pura's dataset (Ooi *et al.*, 2021), are not suitable for evaluating the perceived time of presence. This phenomenon is showcased with a synthetic example on Figure 11.

C. Is AI really necessary ? Why not simply use acoustic indices as source presence predictors ?

While the lack of robustness of acoustic indices was addressed in Section I, we also demonstrated in Section IV that the utilization of PANN-1/3oct enables studying a more extensive number of sound sources. This is achieved without the need to create a specific acoustic index for each distinct sound source, as PANNs inherently feature 527 output sources. This approach opens up possibilities for a more extensive comprehension of the urban acoustic environment, as exemplified with the music, church bells, and civil defense sirens analysis in Figures 7, 8, and 10.

D. Can PANN-1/3oct be used for other applications ? Are all the sound sources predictions reliable ?

We deliberately adopted a straightforward approach to map PANN-1/3oct predictions to the perceived time of presence to demonstrate the full potential of this algorithm. We believe that this methodology can be further refined, and that PANN-1/3oct predictions have the potential to be applied to various other applications.

While our results have demonstrated the predictive capabilities of PANN-1/3oct, extending beyond the primary classes of traffic, voices, and birds, it is important to exercise caution in its application. There is no guarantee of its performance when used on other sound classes for time of presence predictions. Before deploying the model for a new sound class, a preliminary study is essential to verify the model's compatibility and the correlation of its results with the target sound class annotations.

VI. CONCLUSION

In this study, we have explored a novel approach to accurately predict the time of presence of different sound sources in urban soundscapes. Our approach leverages deep learning models, specifically Pre-trained Audio Neural Networks (PANNs), and a transcoding algorithm to convert fast third-octave representations into Mel spectrograms, which are used by PANNs models as input. This method has shown promise in addressing the chal-

lenges of assessing soundscape quality and predicting the perceived time of presence of sound sources.

Our findings demonstrate that PANNs, whether using Mel spectrograms (PANN-Mel) or transcoded Mel-spectrograms from third-octave measurements (PANN-1/3oct), can accurately predict the time of presence of sound sources in various urban environments, specifically for traffic, voices and birds. This approach provides a robust and efficient means of assessing soundscape quality, especially in situations where relying solely on human-generated annotations is impractical. Additionally, we have shown the versatility of such an approach, which allows to potentially assess the presence of over 500 different sound sources. This capability allows for more in-depth analyses of soundscapes, leading to a better understanding of the acoustic environment in urban areas. However, it remains unproven whether those other PANN classes are similarly correlated with time of presence, as the literature on time of presence assessment has focused on only a limited number of sources. In future investigations, we will explore transcoding fast third-octave measurements into audio, enabling the use of any classifier for sound source predictions, and providing the capability to listen to the recordings. This advancement may potentially enrich the applications of third-octave measurements.

In conclusion, our study offers a promising path for advancing our understanding of urban soundscapes, enhancing noise management strategies, and ultimately improving the quality of life for city dwellers. By leveraging artificial intelligence and acoustic measurements, we can gain insights into the intricate relationships between sound sources, human perception, and the quality of urban soundscapes. As the presence of sound sources is closely related to higher-level perceptual attributes such as pleasantness or eventfulness, identifying in-context relationships between sound sources and these perceptual attributes could enable the creation of perception-based visual representations adapted to various stakeholders. Properly displayed, this knowledge can guide decision-makers, city planners, and researchers in creating more pleasant and sustainable sonic environments for everyone.

ACKNOWLEDGMENTS

This work was performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011013544), and has been funded by the AIby4 project (Centrale Nantes and Project ANR-20-THIA-0011).

VII. DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

VIII. REFERENCES

Aletta, F., Kang, J., and Axelsson, O. (**2016**). "Soundscape descriptors and a conceptual framework for developing predictive soundscape models," Landscape and Urban Planning **149**, 65–74 publisher: Elsevier.

Ardouin, J., Baron, J. C., Charpentier, L., Ecotiere, D., Fortin, N., Gontier, F., Guillaume, G., Lagrange, M., Libouban, G., and Picaut, J. (**2021**). "A high density network of low cost acoustic sensors based on wired and airborne transmission of spectral data," in *Euronoise 2021*.

Aumond, P., Can, A., De Coensel, B., Botteldooren, D., Ribeiro, C., and Lavandier, C. (**2017**). "Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context," Acta Acustica united with Acustica **103**(3), 430–443 publisher: S. Hirzel Verlag.

Axelsson, O., Nilsson, M. E., and Berglund, B. (**2010**). "A principal components model of soundscape perception," The Journal of the Acoustical Society of America **128**(5), 2836–2846 publisher: AIP Publishing.

Bansal, A., and Garg, N. K. (**2022**). "Environmental Sound Classification: A descriptive review of the literature," Intelligent Systems with Applications 200115 publisher: Elsevier.

Botteldooren, D., Lavandier, C., Preis, A., Dubois, D., Aspuru, I., Guastavino, C., Brown, L., Nilsson, M., and Andringa, T. C. (**2011**). "Understanding urban and natural soundscapes," in *Forum Acusticum 2011*, European Accoustics Association (EAA), pp. 2047–2052.

Can, A., Picaut, J., Ardouin, J., Crepeaux, P., Bocher, E., Ecotiere, D., and Lagrange, M. (**2021**). "CENSE Project: general overview," in *Euronoise 2021: European Congress on Noise Control Engineering*.

Commission, E. (**2022**). "European Noise Directive 2002/49/EC of the European Parliament and of the Council, of 25 June 2002, relating to the assessment and management of environmental noise (2002)," .

Commission, I. E. *et al.* (**2013**). "Electroacoustics—sound level meters—part 1: Specifications (iec 61672-1)," Geneva, Switzerland .

Farrés, J. C. (**2015**). "Barcelona noise monitoring network," in *Proceedings of the EuroNoise*, pp. 218–220.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (**2017**). "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pp. 776–780.

Gontier, F., Lagrange, M., Aumond, P., Can, A., and Lavandier, C. (**2017**). "An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach," Sensors **17**(12), 2758 publisher: MDPI.

Gontier, F., Lavandier, C., Aumond, P., Lagrange, M., and Petiot, J.-F. (**2019**). "Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques," Acta Acustica united with Acustica **105**(6), 1053–1066 publisher: S. Hirzel Verlag.

Gontier, F., Lostanlen, V., Lagrange, M., Fortin, N., Lavandier, C., and Petiot, J.-F. (**2021**). "Polyphonic training set synthesis improves self-supervised urban sound classification," The Journal of the Acoustical Society of America **149**(6), 4309–4326 publisher: Acoustical Society of America.

Hinton, G., Vinyals, O., and Dean, J. (**2015**). "Distilling the Knowledge in a Neural Network," stat **1050**, 9.

Hong, J. Y., and Jeon, J. Y. (**2015**). "Influence of urban contexts on soundscape perceptions: A structural equation modeling approach," Landscape and Urban Planning **141**, 78–87 publisher: Elsevier.

Hou, Y., Ren, Q., Zhang, H., Mitchell, A., Aletta, F., Kang, J., and Botteldooren, D. (**2023**). "Ai-based soundscape analysis: Jointly identifying sound sources and predicting annoyance," The Journal of the Acoustical Society of America **154**(5), 3145–3157.

ISO, P. (**2014**). "Ts 12913-1, acoustics–soundscape part 1: Definition and conceptual framework.," London, United Kingdom: British Standards Institution .

ISO, P. (**2018**). "Ts 12913-2 12913-1 acoustics—soundscape part 2: Data collection and reporting requirements," London, United

Kingdom: British Standards Institution .
Jeon, J. Y., and Hong, J. Y. (**2015**). "Classification of urban park soundscapes through perceptions of the acoustical environments," Landscape and urban planning **141**, 100–111 publisher: Elsevier.
Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (**2020**). "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 2880–2894 publisher: IEEE.
Lavandier, C., Aumond, P., Can, A., Gontier, F., Lagrange, M., and Petit, G. (**2021**). "Urban sensor network for characterizing the sound environment in Lorient (France) through an automatic assessment of traffic, voice and bird presence ratios," in *European Congress on Noise Control Engineering (EuroNoise)*.
Lavandier, C., and Defréville, B. (**2006**). "The contribution of sound source characteristics in the assessment of urban soundscapes," Acta acustica united with Acustica **92**(6), 912–921 publisher: S. Hirzel Verlag.
Mesaros, A., Heittola, T., and Virtanen, T. (**2018**). "A multi-device dataset for urban acoustic scene classification," in *Proc. Workshop Detection Classification Acoust. Scenes Events*.
Mietlicki, F., Mietlicki, C., and Sineau, M. (**2015**). "An Innovative Approach for long term environmental noise measurement: RUMEUR Network in the Paris Region," in *Proceedings of the EuroNoise*.
Mitchell, A., Oberman, T., Aletta, F., Kachlicka, M., Lionello, M., Erfanian, M., and Kang, J. (**2021**). "Investigating urban soundscapes of the COVID-19 lockdown: A predictive soundscape modeling approach," The Journal of the Acoustical Society of America **150**(6), 4474–4488 publisher: Acoustical Society of America.
Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (**2019**). "Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI" ArXiv:1902.01876 [cs].
Nilsson, M., Botteldooren, D., and De Coensel, B. (**2007**). "Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas," in *Proceedings of the 19th International Congress on Acoustics*.
Ooi, K., Watcharasupat, K. N., Peksi, S., Karnapi, F. A., Ong, Z.-T., Chua, D., Leow, H.-W., Kwok, L.-L., Ng, X.-L., Loh, Z.-A. *et al.* (**2021**). "A strongly-labelled polyphonic dataset of urban sounds with spatiotemporal context," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, pp. 982–988.
Papadakis, N. M., Aletta, F., Kang, J., Oberman, T., Mitchell, A., Aroni, I., and Stavroulakis, G. E. (**2023**). "City, town, village: Potential differences in residents soundscape perception using ISO/TS 12913-2: 2018," Applied Acoustics **213**, 109659 publisher: Elsevier.
Penrose, R. (**1955**). "A generalized inverse for matrices," in *Mathematical proceedings of the Cambridge philosophical society*, Cambridge University Press, Vol. 51, pp. 406–413, issue: 3.
Picaut, J., Can, A., Fortin, N., Ardouin, J., and Lagrange, M. (**2020**). "Low-cost sensors for urban noise monitoring networks—A literature review," Sensors **20**(8), 2256 publisher: MDPI.
Ricciardi, P., Delaitre, P., Lavandier, C., Torchia, F., and Aumond, P. (**2015**). "Sound quality indicators for urban places in Paris cross-validated by Milan data," The Journal of the Acoustical Society of America **138**(4), 2337–2348.
Sueur, J., Aubin, T., Simonis, C., Lellouch, L., Brown, E. C., Depraetere, M., Desjonqueres, C., Fabianek, F., Gasc, A., and LaZerte, S. (**2016**). "Package 'seewave"' .
Tailleur, M., Lagrange, M., Aumond, P., and Tourre, V. (**2023**). "Spectral trancoder: using pretrained urban sound classifiers on undersampled spectral representations," in *8th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.
Tarlao, C., Steffens, J., and Guastavino, C. (**2021**). "Investigating contextual influences on urban soundscape evaluations with structural equation modeling," Building and Environment **188**, 107490 publisher: Elsevier.
Torija, A. J., Ruiz, D. P., and Ramos-Ridao, A. F. (**2013**). "Application of a methodology for categorizing and differentiating urban soundscapes using acoustical descriptors and semantic-differential attributes," The Journal of the Acoustical Society of America **134**(1), 791–802.
Yong Jeon, J., Jik Lee, P., Young Hong, J., and Cabrera, D. (**2011**). "Non-auditory factors affecting urban soundscape evaluation," The Journal of the Acoustical Society of America **130**(6), 3761–3770 publisher: AIP Publishing.