Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques

Félix Gontier [1], Catherine Lavandier [2], Pierre Aumond [2,3]
Mathieu Lagrange [1], Jean-Francois Petiot [1]

[1] LS2N, UMR CNRS 6004, Ecole Centrale de Nantes, F-44321, France
[2] ETIS, UMR CNRS 8051, University of Paris Seine, University of Cergy-Pontoise, ENSEA, CNRS, F-95000, France
[3] IFSTTAR, CEREMA, UMRAE, F-44344, Bouguenais, France

Abstract

The impact of urban sound on human beings has often been studied from a negative point of view (noise pollution). In the two last decades, the interest of studying its positive impact has been revealed with the soundscape approach (resourcing spaces). The literature shows that the recognition of sources plays a great role in the way humans are affected by sound environments. There is thus a need for characterizing urban acoustic environments not only with sound pressure measurements but also with source-specific attributes such as their perceived time of presence, dominance or volume.

This paper demonstrates, on a controlled dataset, that machine learning techniques based on state of the art neural architectures can predict the perceived time of presence of several sound sources at a sufficient accuracy. To validate this assertion, a corpus of simulated sound scenes is first designed. Perceptual attributes corresponding to those stimuli are gathered through a listening experiment. From the contributions of the individual sound sources available for the simulated corpus, a physical indicator approximating the perceived time of presence of sources is computed and used to train and evaluate a multi-label source detection model. This model predicts the presence of simultaneously active sources from fast third octave spectra, allowing the estimation of perceptual attributes such as pleasantness in urban sound environments at a sufficient degree of precision.

PACS numbers: 43.66.Lj (Perceptual effects of sound), 43.60.-c (Acoustic signal processing)

1 Introduction

Leveraging the advent of the Internet of Things (IoT) and the availability of low-cost sensor networks [1, 2] could allow us to characterize the sound environment in a way that is closer to the perception of city dwellers [3]. This requires the identification of its composition in terms of sources of interest from acoustic measurements at a reasonable cost. Most current monitoring applications rely on the measurement of sound levels on time scales from 1s to several hours [4, 5, 6], which offers limited information regarding the content of the sound environment. Conversely, sound environment recognition and event detection applications operate on spectral representations of recorded audio such as Mel spectrograms or Mel frequency cepstrum coefficients, on much finer time scales in tens of milliseconds [7, 8, 9]. Using such information-rich representations of the signal also allows the computation of commonly used monitoring acoustical indicators. Though, it may not be desirable in long-term monitoring applications where analysis is performed offlinen as it requires large transmission bandwidth and storage capabilities, and raises privacy concerns if intelligible speech can be retrieved [10]. Some monitoring applications [11, 12, 13] use third octave sound level fast (125ms) measurements to underline relevant spectral information. This representation is easier to store in the long term and reduces privacy concerns [10]. In [11], the authors successfully linked acoustical indicators derived from third octave sound levels to the perceived activity of specific sound sources in recordings of polyphonic sound scenes. Though the discriminative properties of the proposed indicators are limited and the underlying methodology cannot be easily extended to larger sound source taxonomies, it demonstrates that the information content of third octave spectra is sufficient to identify sound sources.

The characterization of urban soundscapes through standardized perceptual descriptors has been extensively studied [14, 15, 16, 17, 18]. A two-dimensional soundscape model of perceived quality emerges where one dimension corresponds to the pleasantness of the soundscape, and the orthogonal one corresponds to its enventfulness [15, 11, 19]. In the case of pedestrians, three major source types are found to contribute to these dimensions in urban contexts: technologi-

cal, human and natural sources [13, 15], though the exact taxonomy of sources used differs across studies [20, 21, 22]. Furthermore, traffic noise, human voices and birds calls can respectively be used as a proxy of these source types [23, 24, 11]. In existing models, source activity can be quantified by the dominance [25, 15], the time of presence [24], or the volume. The dominance is built on coupled notions of time and volume, while the time of presence is more easily linked to the physical activity of sources. In a more holistic approach of soundscape quality evaluation, the time of presence should thus be easier to identify from acoustic monitoring measurements and would provide more complete information on the content of polyphonic urban environments [26].

In this paper, we do so by developing new indicators relying on source recognition models based on deep learning techniques, which has demonstrated state-of-the-art performance in many tasks studied in the machine listening community [27]. In the context of urban sound environments, machine listening has been successfully applied to sound event detection [28, 29, 30, 31], sound scene classification [12, 32, 33, 34, 35] and soundscape quality evaluation [7, 36, 31]. A wide range of architectures exist, the most common of which are convolutional and recurrent neural networks [37]. Large amounts of data with task-specific annotations are required for deep learning architectures to learn to extract relevant information from recordings [38, 27, 39]. To the best of our knowledge such databases do not currently exist in the literature in the case of source recognition with labels consisting of perceived source activity. It is possible to manually annotate each recording in the corpus. However, this process would be time-consuming, and must be repeated to extend the taxonomy of relevant sound sources. As an alternative, we consider the use of simulated corpora [40, 41], of which the generation procedure provides complete knowledge about the composition of sound scenes. By using a simple indicator computed on this information-rich data to approximate the source-specific perceived time of presence labels, the excerpts composing large corpora may be automatically annotated without the need for additional subjective inputs. Thanks to this data generation procedure, we are able to validate the proposed approach.

More precisely, the contributions of this paper are three-fold:

1. Provide a corpus of simulated scenes for which relevant acoustic properties are well controlled and perceptual judgements are available[1].

2. Propose numerical means[2] for predicting the perceived time of presence of sound sources from raw acoustic data using deep learning approaches, with a larger corpus of simulated scenes[3].

3. Demonstrate that the proposed method can be applied to the prediction of a perceptual descriptor of soundscapes (pleasantness) to a sufficient degree of accuracy through a comparative study with state of the art approaches.

Section 2 presents the generation procedure for a large corpus of simulated sound scenes, its validation through a listening test in the context of a study on perceived soundscape quality, and a study of available acoustical indicators for the automatic annotation of source activity. In Section 3 a deep learning architecture that performs multi-label sound source recognition at relevant time scales is trained and evaluated. Section 4 then presents an application of this model to the prediction of pleasantness in urban environments.

2 Corpus

2.1 Generation procedure

Considering machine learning techniques to detect the presence of sources requires the availability of annotated data. Thus, in order to train deep learning architectures for source recognition, a large corpus is constructed. This corpus is composed of sound scenes simulated using the *simScene* software, an open source software library in Matlab that allows the simulation of sound scenes as the additive composition of sound sources, using a database of recordings associated to isolated occurences of these sources[4]. Here we choose to restrict the taxonomy of active sources to *traffic*, *human voices* and *birds*, as these sources are found to be the most influent on soundscape quality for pedestrians [23]. The corresponding isolated samples database is constructed from excerpts of the LibriSpeech [42] corpus for voices and Freesound[5] contributions for remaining sources. *simScene* generates original scenarios using the following information:

- The probability of appearance of a given source in the scene, for which events and backgrounds are considered separately,
- Event-to-background ratios in dB for all events and backgrounds compared to a main background source, drawn from gaussian distributions,
- The inter-onset of event occurrences in seconds, also drawn from gaussian distributions.

[1] Corpus available at `https://zenodo.org/record/3248734#.XQjC4v7gqUk`

[2] Open source code is available at `https://github.com/felixgontier/soundSourcePresenceEstimation`.

[3] Corpus available at `https://zenodo.org/record/3248703#.XQjDVv7gqUk`

[4] `https://bitbucket.org/mlagrange/simscene`

[5] `https://freesound.org`

Scenarios generated by *simScene* should cover most real-life situations while remaining perceptually plausible. To guarantee the realism of created sound scenes, the information needed for *simScene* generation is based on a corpus of 74 recordings from the GRAFIC project [11]. These 74 sound scenes are obtained during 4 soundwalks in 19 locations in the 13th district of Paris, France, and range from 55 s to 4.5 m in duration. In [43], the extracts are manually annotated and classified in terms of ambiance (*park*, *quiet street*, *noisy street* and *very noisy street*). The available annotations include:

- Background sources that are present throughout the whole scene, and their respective sound level considered constant,
- Sound events characterized for each occurrence by their source type, onset-offset and event-to-background ratio in dB.

This information is reused to extract statistics on background and event sources activity, which are then used as inputs by *simScene*. Though, adjustments are made on the variance of the considered properties to extend the range of covered scenarios. Furthermore, since voice events in the isolated samples database consist mostly of read English recordings, the voice event mean event-to-background ratios are reduced for all ambiances to improve realism. An additional *square* ambiance is added with properties derived empirically from available data, with predominant voice activity. Diverse new scenarios are then generated by sampling the background and event properties distributions and used with samples randomly selected from the database of isolated occurrences to simulate sound scenes. At the end of the simulation process the sound level of each sound scene is randomly sampled and conditioned to the ambiance according to typical sound levels in urban context.

The simulated deep learning corpus is composed of two subsets:

- The development set, made of 400 scenes of 45 s each (total duration 5 hours), which is used during the training process,
- The evaluation set, made of 200 scenes of 45 s each (total of 2.5 hours), which is used to compute the performance of the trained model and its generalization capabilities.

The generation procedure is the same for both datasets. Though, it is important in deep learning applications to ensure a complete independance of the training and evaluation sets. To do so, the isolated samples database from which *simScene* assemble extracts to generate a sound scene is split in the same proportions as the two datasets: two-thirds for the development set and the remaining one-third for the evaluation set. As a result, the two corpora differ by the scenarios and isolated samples used in their generation.

2.2 Perceptual experiment

A listening test is conducted in order to validate the perceptual correspondence of the proposed corpus with urban sound scene recordings, as well as to obtain subjective annotations for the design and evaluation of the time of presence estimation model. Given the aim of this test, a separate corpus is constructed. This corpus contains 100 sound scenes, including 75 simulated scenes as well as 6 recorded and 19 replicated scenes from the GRAFIC project.

First, 200 simulated scenes of 45s each with diverse new scenarios are generated using the procedure described in Section 2.1, though with a third isolated samples database to ensure independance from the deep learning datasets. These scenes are equally distributed among the five considered ambiances (resp. *park*, *quiet street*, *noisy street*, *very noisy street* and *square*). For each scene the perceived time of presence for the traffic, voice and bird sources is estimated using the indicators proposed in [44]. Ideally, sound scenes should cover the resulting 3-dimensional space in a homogeneous way. To do so, the 75 scenes that maximize the minimum pairwise distance in the 3-dimensional space are selected. Figure 1 shows the distribution of the 75 selected extracts. Playback sound levels are realistically drawn for each scene as described in Section 2.1, and range from 46.6 dB SPL to 77.1 dB SPL over the 75 simulated scenes.

In order to validate the generation procedure of new scenarios, the listening test corpus is extended with replications of recorded sound scenes. Reference recordings are obtained from one of the four soundwalks performed in [11], including 19 locations (noted P1-19) with diverse environments. For each of the 19 corresponding recordings, 45 seconds of audio in a single channel are extracted. The 45 s segments are selected to represent the properties of their respective ambiances in terms of source composition, without single events overwhelming their overall perception. The manual annotations available in [43] of background and event information are then used to replicate the sound scenes using *simScene*. Furthermore, the original 45 s recordings from 6 locations (P1, P3, P4, P8, P15 and P18) are added to the experiment corpus to evaluate changes in perception yielded by the replication process, as they explore diverse real-life situations with respect to ambiance categorization. The 6 recorded and 19 replicated scenes are normalized so that their playback sound level through the restitution system is the same as measured during recording, with a range from 63.9 dB SPL to 79.4 dB SPL.

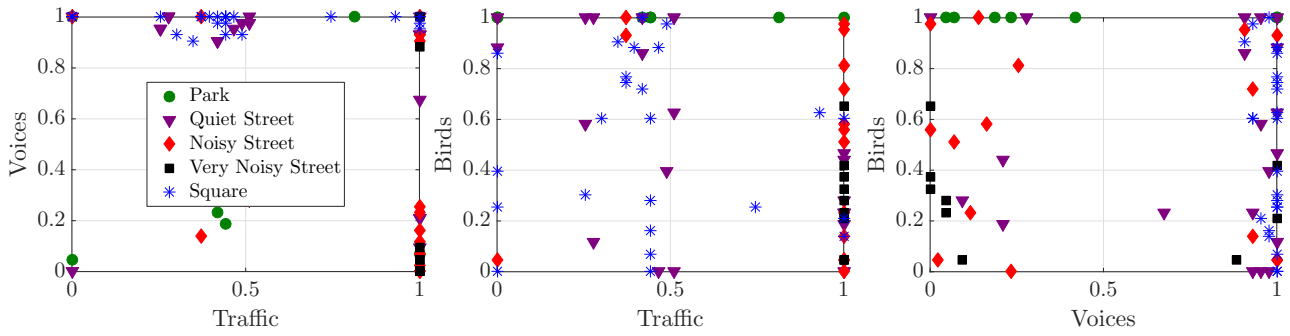

Figure 1: Physical estimations [44] of the perceived time of presence of traffic, voices and birds sources for the corpus of simulated scenes with scenarios generated by *simScene*.

During the test, participants are asked to evaluate sound scenes using 8 criteria represented by 11 point scales (0-10). These scales are presented in French and translated in this report using standard terminology. The first 5 scales relate to general properties of the scene:

- Pleasantness: *Unpleasant - Pleasant* (*Désagréable - Agréable*) - P,
- Liveliness: *Inert, amorphous - Lively, eventful* (*Inerte, amorphe - Animé, mouvementé*) - L,
- Overall loudness: *Quiet - Noisy* (*Silencieux - Bruyant*) - OL,
- Interest: *Boring, uninteresting - Stimulating, interesting* (*Ennuyeux, inintéressant - Stimulant, intéressant*) - I,
- Calmness: *Agitated, chaotic - Calm, tranquil* (*Agité, chaotique - Calme, tranquille*) - C.

These quantities are typically studied in the perceptual characterization of sound scenes [15, 11, 13]. Additionally, to assess the perceived source activity 3 questions are presented to the participants and evaluated on the same 11 point scales:

- Time of presence of traffic, voice and bird sources: *Never - Continuously* (*Jamais - Continuellement*) - resp. $T_{T,p}$, $T_{V,p}$ and $T_{B,p}$

where p denotes a perceptual evaluation.

Prior to the test, a short verbal introduction is given to the participants and the interface is introduced to ensure that the quantities are well understood. Although the corpus is comprised of 100 sound scenes, participants only evaluate 50 scenes: all listen to the 6 recorded and 19 replicated sound scenes, then to 25 of the 75 simulated with new scenarios according to a balanced incomplete block design [45]. The selection of simulated scenes is done so that all scenes in the sub-corpus are evaluated by the same number of participants. All participants are first presented with the most quiet then loudest of the recorded scenes (resp. P3 and P15). A random listening order is generated for each subject to control ordering effects for the remaining of the test. Participants can listen to each scene once, and have to listen to the full extract and to answer all questions before being allowed to proceed to the next scene.

The scenes are played at a given sound level as discussed in Section 2.1, through the same computer and sound card configurations. Beyerdynamics DT-990 Pro headphones are used by all participants. The calibration of the headphone was carried out in a free field situation (pink noise through a Genelec 1031A loudspeaker in front of the head) and consisted in characterizing the relationship between voltage at the headphone terminal and the sound pressure at the entrance of a blocked ear canal. To do so, small DPA 4060 microphones have been set at the entrance of the ear canals of a human subject [46]. From this information a scaling factor is applied to the sound scenes to ensure that they are heard at the desired sound level by every listener.

A total of 23 students aged from 22 to 23 years including 16 males and 7 females at Ecole Centrale de Nantes completed the test, all reported normal hearing (note that with 23 subjects, the incomplete block design is not perfectly balanced). All participants gave written consent prior to the experiment, and evaluations were further anonymized.

2.3 Simulated corpus validation

The effect of the *simScene* generation procedure on perception is first investigated. To do so, perceptual responses are compared for the 6 recorded scenes and 6 corresponding replicated scenes, which share common scenarios. Table 1 shows the mean differences between assessments for pairs of scenes with equivalent scenarios. Wilcoxon signed-rank tests [47] are implemented for each scene and perceptual scale to outline significant differences between assessment distributions, which are shown in bold. As the data is discretely distributed, zero differences between paired samples are included using Pratt's modification of the test [48]. On the first five scales, all mean differences

are lower than 2 points on the 11-point Likert scale. Though, significant differences are outlined that can be linked to corresponding discrepancies in source-specific parameters. The highest difference (-5.22) is found for the assessment of the time of presence of traffic in the location P4. For this location, the background traffic in the recorded scene varies along time, it is louder in the first half of the scene than in the second half. Replicating this scene using simScene imposes a constant sound level for background sources. Thus, the background traffic is louder in the replicated scene than it is in the recording for about half of its duration. To a lesser extent the same issue explains the large difference (-4) in the assessed time of presence of voices in the location P3. Discrepancies on source-specific scales can also be interpreted by the choice of isolated samples, which is semi-random and based on a high-level source taxonomy. For example, no difference is made during annotation between child or adult speech, or depending on its expressiveness. Though, overall no consistent difference between the perception of recorded and replicated scenes emerges for the studied points.

Next, the perceptual space generated by the experiment's five general scales (pleasantness, liveliness, overall loudness, interest and calmness) is studied to validate the use of simulated sound scenes with new scenarios as well as reduced source complexity. It is obtained by performing a principal components analysis on the corresponding perceptual responses averaged along participants. No standardization is applied to the data. Figure 2 and Figure 3 compare the results for scenes based on recordings (n=25) and new scenarios (n=75) respectively. The resulting spaces are similar, with only overall loudness and pleasantness axes slightly rotated between the two subcorpora. For both sets the variance explained by the first two components is similar, resp. 79.4% - 18.1% and 79.6% - 15.2%. Furthermore, these representations are comparable to those found in previous work on perceptual dimensions [15, 16]. Thus, the use of simulated scenes based on both real and new scenarios does not result in major differences in the relations between perceptual quantities. Additionally, the assessments averaged on all subjects for active individuals (simulated scenes) are projected onto the principal components space for and represented in Figure 3 as dots. The assessments for recorded and replicated scenes are then projected as supplementary individuals on this space and represented as crosses. These projections show that the space covered by scenes based on new scenarios covers that of the studied real-life environments. This further demonstrates the diversity of scenarios created by the scene generation procedure.

Discrepancies between projections of original and replicated scenes are highlighted in Figure 2 using arrows. The standard deviation of assessments is represented using ellipses for one location (P1). The projections of assessment distributions consistently overlap for all pairs of recorded and replicated scenes[6]. This illustrates the results in Table 1 where perceptual assessments were not found to differ significantly overall.

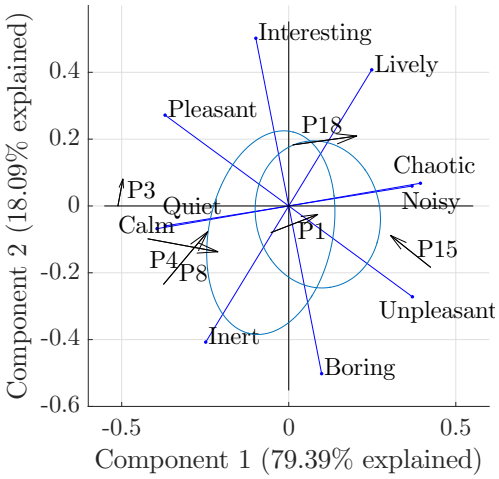

Figure 2: Biplot of the principal components analysis of average assessments for the 5 general questions on the 6 recorded and 19 replicated scenes (n=25). Arrows show discrepancies between corresponding recorded and replicated scenes. For the P1 recording location the standard deviation of the projection of individual responses is shown using ellipses.

2.4 Acoustical indicators for soundscape description

To train and evaluate a deep learning architecture for sound source recognition using the corpus described in Section 2.1, the activity of sound sources present in each scene must be known. Though, the manual annotation of such information is impractical: it is extremely time-consuming due to the size of the corpus, and the process must be repeated for each studied sound source. Several indicators are identified in the literature that correlate well with perceptual parameters [11, 44, 24]. These indicators may be useful to automatically annotate the deep learning corpus without the need for additional human input.

First, some indicators can be computed directly from the mixed audio. Typically this includes indicators derived from sound level measurements used in monitoring applications. For this study the following variables are considered, and computed with a time frame of 1 s using the Matlab ITA-toolbox [49]:

- Z-weighted L_{eq} and A-weighted LA_{eq} equivalent sound levels in dB and dBA respectively.
- L_{10}, L_{50} and L_{90}: 10th, 50th and 90th percentiles of the Z-weighted sound level. The L_{10} is often

[6] Data available here `http://felixgontier.github.io/soundSourcePresenceEstimation/web/index.html`

Table 1: Mean differences of perceptual assessments between original and replicated sound scenes. Significant differences as per a Wilcoxon signed-rank test are shown in bold (n=23, p<0.05)

	P	L	OL	I	C	$L_{T,p}$	$T_{T,p}$	$T_{V,p}$	$T_{B,p}$
P1	0.43	**-1.65**	**-1.04**	0.43	0.13	**-0.91**	0.39	**-2.09**	0.61
P3	0.26	-0.43	0.30	-1	0.30	0.35	1.04	**-4**	0.22
P4	0.91	0	**-1.83**	0.48	**1.30**	**-0.96**	**-5.22**	**1.43**	0.04
P8	0.26	**-1.65**	-0.87	-0.96	0.65	**-2.04**	**-0.91**	0.09	**-1.43**
P15	**-1.35**	0.52	0.52	**-1.17**	0.09	0.61	0.13	**1.96**	**-2.74**
P18	**1.13**	-0.30	**-1.17**	-0.43	**1.39**	-1.04	**-1.83**	**0.83**	**1.30**

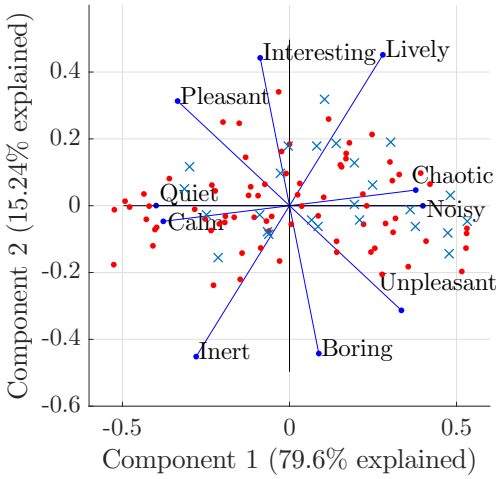

Figure 3: Biplot of the principal components analysis of average assessments for the 5 general questions on the 75 simulated scenes (n=75). Assessments of simulated scenes (active individuals) are projected as dots, and recorded and replicated scenes (supplementary individuals) are projected as crosses.

associated to events and the L_{90} to background activity, while the L_{50} is a measurement of the overall sound level.

- LA_{50}: 50th percentile of the A-weighted sound level, with similar properties as the L_{50}.
- $L_{50,1kHz}$: 50th percentile of the Z-weighted sound level for the 1kHz frequency band, also a good descriptor of the overall sound level of the scene.
- $LA_{10} - LA_{90}$: Emergence indicator included in the pleasantness model presented in [24].

The time and frequency second derivative ($TFSD$) is introduced in [11] as a descriptor of perceived source activity. Its expression is:

$$TFSD_{f,t} = \frac{|\frac{d^2L(f,t)}{df\,dt}|}{\sum_{f_1=16kHz}^{f_1=31.5Hz}|\frac{d^2L(f_1,t)}{df_1\,dt}|} \quad (1)$$

where $L(f,t)$ is the third-octave spectrum of the signal. This indicator represents the variations in both the time and frequency dimensions to highlight sources of interest. For example, bird activity is characterized by narrow-band energy with fast paced variations in time, which translates into high $TFSD$ values in the corresponding frequency range. The $TFSD$ is computed for the 4kHz band and 125ms measurements ($TFSD_{4kHz(1/8s)}$), and for the 500Hz band with 1s measurements ($TFSD_{500Hz,1s}$) as estimates of the perceived activity of birds and voices respectively.

In the case of simulated scenes, the generation process outputs ground truth source contributions as separate channels. This information could be used to compute additional indicators that describe source activity perception with better accuracy. The following are computed for traffic, voice and bird sources: the equivalent sound level $L_{eq,s}$ for source s and the source emergence ΔL_s, taken as the difference between the equivalent sound level of source s and that of all other sources combined. Events and background occurences of the same source are added in this study. Next, the $\hat{T}_s(\alpha,\beta)$ time of presence approximation proposed in [44] is considered. $\hat{T}_s(\alpha,\beta)$ is based on a binary source emergence model computed on the third-octave band emergence spectrum $\Delta L_s(t,f)$. It is parametrized by α and β thresholds:

$$\hat{T}_s(\alpha,\beta) = \frac{1}{N_t}\sum_{t=1}^{N_t}\mathbb{1}\left[\frac{\sum_{f=1}^{N_f}\Delta L_s(t,f)\mathbb{1}_{\Delta L_s(t,f)>\alpha}}{\sum_{f=1}^{N_f}\mathbb{1}_{\Delta L_s(t,f)>\alpha}} > \beta\right] \quad (2)$$

$$\alpha_{opt},\beta_{opt} = \arg\max_{\alpha,\beta}\frac{1}{N_s}\sum_{s=1}^{N_s} r\left(T_{s,p},\hat{T}_s(\alpha,\beta)\right) \quad (3)$$

where r is the Pearson's correlation coefficient, s denotes the sound source, N_s is the number of sources in the taxonomy and equals 3 in this study, t is the time frame and $T_{s,p}$ corresponds to the perceived time of presence assessments averaged per scene. This indicator evaluates the presence or absence of a given source in a small time frame. First, frequency bands for which the source is emergent by more than the α threshold value are isolated. Then, the source is considered present if the emergence for these bands is on average greater than the β threshold. A time of presence estimation is obtained for each source by

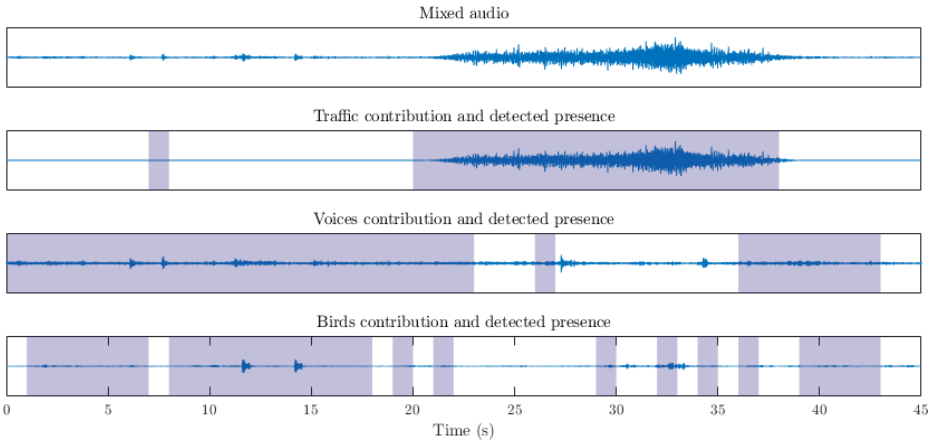

Figure 4: Example of simulated sound scene with ground truth source contributions and resulting presence predictions by the deep learning architecture (greyed out areas)

averaging along the N_t time frames composing the scene. The optimal threshold values are optimized via grid search so that the time of presence estimation $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ is best correlated with the average perceived time of presence $T_{s,p}$ as obtained by human evaluation. Both thresholds are optimized once in this study using data from the listening test in Section 2.2 and equal $\alpha_{opt} = -14dB$ and $\beta_{opt} = -7dB$, though several other pairs of values yielded close performance as per the optimization metric defined in eq.2.

In the remainder of this paper, the subscript s is replaced with the corresponding source initial: T for traffic, V for voices and B for birds.

The considered indicators are compared to subjective assessments obtained during the listening test to identify their capacity to explain dimensions of soundscape perception. The analysis is performed using the arithmetic mean of subjective assessments over all participants for replicated and simulated scenes. The source-specific sound level $L_{eq,s}$, emergence ΔL_s and estimated time of presence $\hat{T}_s(\alpha, \beta)$ indicators cannot be directly computed on the 6 recorded scenes, as their ground truth source composition is unknown. Thus, these 6 scenes are excluded of the study. Two simulated scenes contained only one source, leading to infinite source emergences. Statistics including these scenes cannot be computed and they are thus excluded in this analysis, which as a result includes $n = 92$ scenes.

Table 2 shows the Pearson's correlation coefficients between computed physical indicators and perceptual assessments. First, all global sound level indicators correlate well (r>0.9) with the perceived overall loudness. Regarding source-specific perceptual parameters, the $L_{eq,s}$ correlates consistently well with the $T_{s,p}$ of corresponding source (r=0.71). Conversely, the emergence ΔL_s fails to represent the perceived bird activity, and correlations are weak for other sources. The proposed estimates of the time of presence $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ show strong correlations to their perceived counterparts (r>0.8), though this is expected as α_{opt} and β_{opt} are optimized to this aim. They also display good source discrimination properties, as no significant correlation is found between voices and birds, and perceptual assessments of traffic were already correlated with those of other sources in Table 4. They are also better predictors than the $TFSD_{500Hz,1s}$ and $TFSD_{4kHz,1/8s}$ indicators for voice and bird activity in this corpus. Thus, because source contributions are available in simulated scenes they can be automatically annotated in terms of estimated perceived time of presence using the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ indicator for the three sources.

3 Prediction of time of presence of sources

3.1 Deep learning for presence prediction

A deep learning model is implemented using the Python Pytorch [50] framework and trained on the corpus presented in Section 2.1 for source time of presence prediction.

The developped model should extract relevant source information from a representation of the audio signal. Typically, spectral representations are preferred to the raw audio waveform because of the regularities they underline in the signal. Here, the third-octave spectrum is considered as it is commonly used in acoustic monitoring applications [1, 10]. Third-octave spectra are computed for 125 ms frames and 29 frequency bands in the $20\ Hz - 12.5\ kHz$ range as the input signal representation. Instead of a regression task where the output is directly the time of presence, a multiple label classification task on 1 s texture frames is preferred as the resulting training procedure is easier. Individual inputs are thus obtained by splitting the resulting spectrograms into texture windows

Table 2: Pearson's correlation coefficients between physical and perceptual indicators (n = 92, *: p<0.05, **: p<0.01). Non significant correlations at the 5% threshold are noted NS.

	P	L	OL	I	C	$L_{T,p}$	$T_{T,p}$	$T_{V,p}$	$T_{B,p}$
LA_{eq}	-0.86**	0.68**	0.92**	-0.37**	-0.88**	0.77**	0.66**	NS	-0.41**
LA_{50}	-0.84**	0.67**	0.91**	-0.33**	-0.87**	0.71**	0.63**	NS	-0.35**
L_{eq}	-0.88**	0.67**	0.91**	-0.44**	-0.88**	0.83**	0.71**	NS	-0.46**
L_{10}	-0.87**	0.65**	0.90**	-0.44**	-0.86**	0.84**	0.71**	NS	-0.47**
L_{50}	-0.89**	0.65**	0.92**	-0.43**	-0.89**	0.77**	0.71**	NS	-0.44**
L_{90}	-0.86**	0.68**	0.92**	-0.39**	-0.89**	0.71**	0.67**	NS	-0.40**
$L_{50,1kHz}$	-0.88**	0.69**	0.92**	-0.42**	-0.89**	0.74**	0.73**	NS	-0.50**
$L_{10} - L_{90}$	NS	NS	-0.24*	NS	-0.22*	NS	NS	NS	NS
$TFSD_{500Hz,1s}$	NS	0.41**	NS	0.28**	NS	-0.24*	-0.39**	0.74**	NS
$TFSD_{4kHz,1/8s}$	0.52**	-0.43**	-0.49**	0.41**	0.52**	-0.45**	-0.54**	NS	0.63**
$L_{eq,T}$	-0.58**	NS	0.46**	-0.46**	-0.42**	0.77**	0.71**	NS	-0.36**
$L_{eq,V}$	NS	0.50**	0.31**	NS	-0.37**	NS	NS	0.71**	-0.40**
$L_{eq,B}$	0.27*	NS	NS	0.35**	NS	-0.25*	-0.24*	NS	0.71**
ΔL_T	-0.45**	NS	0.26*	-0.59**	-0.22*	0.63**	0.66**	-0.51**	-0.26*
ΔL_V	NS	0.50**	NS	0.35**	NS	-0.27**	-0.38**	0.59**	NS
ΔL_B	0.21*	-0.25*	-0.26*	NS	0.25*	-0.24*	-0.25*	NS	NS
$\hat{T}_T(\alpha_{opt}, \beta_{opt})$	-0.53**	NS	0.35**	-0.57**	-0.29**	0.56**	0.81**	-0.39**	-0.37**
$\hat{T}_V(\alpha_{opt}, \beta_{opt})$	NS	0.44**	NS	0.35**	NS	NS	-0.39**	0.81**	NS
$\hat{T}_B(\alpha_{opt}, \beta_{opt})$	0.56**	-0.30**	-0.46**	0.55**	0.51**	-0.46**	-0.57**	NS	0.91**

of 1 s duration. The spectral blocks of dimension 29x8 are then processed independently by the model.

Ground truth target outputs are associated to each input frame to train the model. Considering the results discussed in Section 2.4, the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ time of presence estimation seems well suited to automatically label the dataset for this task. Binary presence values on 1 s frames obtained during the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ computation before averaging in eq.2 are thus used as individual weak labels.

The architecture of the model is shown in Figure 5. The model includes 4 blocks of convolutional layers followed by leaky rectified linear unit (LeakyReLU) activations of expression $y = max(0.1x, x)$. The convolutional layers have respectively 128, 64, 32 and 32 output channels, and a common kernel size of 5x5. The output of the last block is flattened then goes through a fully connected layer with output size 3. A final sigmoid activation is used in order to obtain outputs in the 0-1 range, which correspond to the presence of traffic, voices and birds in the 1 s frame respectively. During training these values are directly compared to presence labels given by $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ using a binary cross-entropy cost function:

$$BCE(y, \hat{y}) = -\sum_s y_s log\,(\hat{y}_s) + (1 - y_s) log\,(1 - \hat{y}_s) \quad (4)$$

where s is the source, y_s and $\hat{y}_s$ are the target and predicted presence for source s in the 0-1 range. This loss function is minimized using the Adam algorithm [51] on batches of 1 s examples. During evaluation, a threshold of 0.5 is independently applied to the 3 outputs to obtain a binary presence value for each source: each source is considered absent when the model outputs a value lower than 0.5 and present when it outputs 0.5 or higher. The time of presence estimation is then obtained by averaging presence labels of all 1 s time frames corresponding to the same scene.

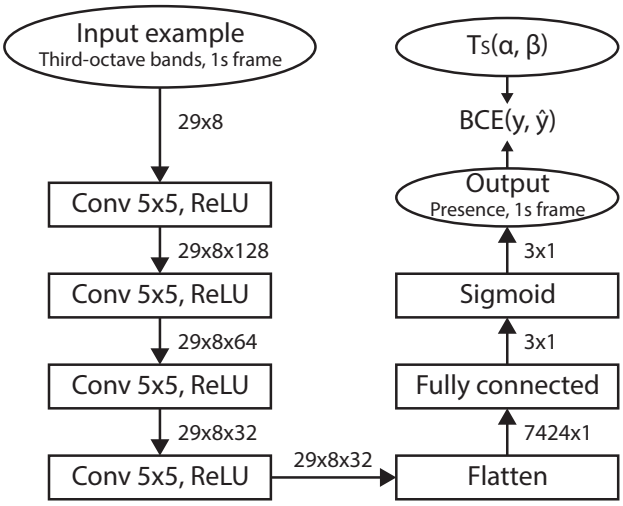

Figure 5: Architecture of the deep learning model used for source presence prediction.

3.2 Results

The performance of the source detection deep learning architecture presented in Section 3.1 on the evaluation dataset is summarized in Table 3. The overall

presence detection accuracy is 92.11%. The model performs similarly well for the three sources, ranging from 90.08% for birds to 94.09% for traffic. Regarding the type of errors, they are equally split between false positives and false negatives. Though, these rates vary depending on the type of source. Traffic has a very low false positive rate at 2.46% and high false negative rate, while birds display the highest false positive rate at 12.30%. This is expected given the spectral content related to these sources. Traffic is usually located in lower frequency bands than voice or bird events, but may contain high frequency components mistaken for birds by the model. The resulting root mean squared error on the time of presence is 12% overall, and is close for the three sources.

Figure 4 presents an example of predictions for a sound scene in the listening test corpus. The model takes the mixed audio (top) as its input, and outputs binary presence predictions for traffic, voice and bird sources on 1s time frames. Here the ground truth contributions are shown as reference, though they are never available to the deep learning architecture. A traffic event masks voice and bird activity from 23s to 37s, and an obvious false positive for traffic presence is visible at 7-8s.

4 Application to pleasantness prediction

In the context of urban soundscape quality assessment for pedestrians, pleasantness emerges as the first affective dimension [15]. Pleasantness has been modeled from both the perceived overall loudness in a scene [52, 53] and its content in terms of active sources [15, 23, 24, 11]. To demonstrate the potential of the proposed method for predicting the time of presence of sources, it is applied to pleasantness estimation.

4.1 Baseline models of pleasantness

Baseline models of pleasantness are constructed using data from the experiment in Section 2.2. We first study the relationships between perceptual scales with respect to existing work. Table 4 shows the Pearson's correlation coefficients between pairs of parameters with assessments averaged for each scene (n=100). The correlations between perceived source activity and pleasantness are consistent with the literature [11, 44]. Pleasantness (P) is mainly influenced negatively by overall loudness (OL) and traffic and positively by birds. In previous studies a small positive contribution of voices to pleasantness was found, while no direct relation is visible from the data gathered in this study. This can be explained by the choice of speech samples used during generation, which consist of read audiobooks extracts and thus may sound unnatural in the considered urban environments. Relations between source-specific parameters are weak with the exception of traffic being negatively correlated with birds (r=-0.42).

Multilinear regression models of pleasantness are built as a function of the overall loudness and the source-specific time of presence of traffic, voices and birds, with assessments averaged for each scene (n=100). All possible combinations of variables among the 4 predictors (15 combinations) are considered, and to ensure that no multi-collinearity exists between predictors, a variance inflation factor (VIF) test is performed prior to each regression. Only combinations for which all predictors verify $VIF < 5$ are considered valid. The best resulting model in terms of R^2_{adj} with statistically significant coefficient estimates ($p < 0.05$) is:

$$\hat{P}_{1,p} = 8.99 - 0.67\ OL - 0.15\ T_{T,p} + 0.08\ T_{V,p} + 0.12\ T_{B,p} \quad (5)$$

where OL, $T_{T,p}$, $T_{V,p}$ and $T_{B,p}$ represent the perceived overall loudness and time of presence of traffic, voices and birds respectively. This expression is close to existing models of pleasantness in the literature [24, 11], with both the overall loudness and traffic activity negatively contributing and bird activity positively contributing to pleasantness. A small positive contribution of the time of presence of voices is also found on this corpus despite no significant correlation underlined in Table 4.

Similarly, multilinear regression models of pleasantness are constructed using acoustical indicators described in Section 2.4. Namely the L_{50} and $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ indicators are used as predictors, as they display the highest correlation values with the overall loudness and time of presence of sources respectively. Again, all combinations of the proposed physical variables are considered, and a variance inflation factor check ($VIF < 5$) is performed on predictors to ensure that no multi-collinearity between predictors exists in a model. On the present corpus, the best model in terms of R^2_{adj} is:

$$\hat{P}_{1,\varphi} = 16.74 - 0.18\ L_{50} + 1.01\ \hat{T}_B(\alpha_{opt}, \beta_{opt}) \quad (6)$$

where φ indicates a model from physical variables. Indicators related to traffic and voices activity are absent compared to the perceptual model. This is consistent with the results underlined in Table 2: the L_{50} used as a predictor is correlated to the traffic time of presence $T_{T,p}$ that appeared in the perceptual model $\hat{P}_{1,p}$ (eq.5), and no strong contribution of objective variables representative of voices to pleasantness is identified in this corpus. These results can be attributed to the diversity of the studied corpus as only traffic, voice and bird sources are present in

Table 3: Performance of the model predicting source presence with binary ground truth labels obtained from $\hat{T}_s(\alpha_{opt}, \beta_{opt})$. Presence metrics are computed for n=8600 1s frames and time of presence metrics on n=200 45s scenes. (TP: true positive, TN: true negative, FP: false positive, FN: false negative)

	All sources	Traffic	Voices	Birds
Presence accuracy (%)	92.11	94.09	92.15	90.08
Presence TP (%)	91.96	92.70	89.93	93.37
Presence TN (%)	92.30	97.54	94.76	87.70
Presence FP (%)	7.70	2.46	5.24	12.30
Presence FN (%)	8.04	7.30	10.17	6.63
$\hat{T}_s(\alpha_{opt}, \beta_{opt})$ RMSE	0.12	0.13	0.10	0.11

Table 4: Pearson's correlation coefficients between perceptual scales at the scene level (n=100, *: p<0.05, **: p<0.01)

	P	OL	$T_{T,p}$	$T_{V,p}$	$T_{B,p}$
P	1	-0.89**	-0.76**	0.05	0.57**
OL		1	0.59**	0.17	-0.45**
$T_{T,p}$			1	-0.35**	-0.42**
$T_{V,p}$				1	-0.21*
$T_{B,p}$					1

simulated sound scenes. As a result global sound level measurements tend to be correlated with the presence of traffic in the simulation process. Quieter environments such as parks and quiet streets are less likely to have continuous traffic, while high sound levels in busy streets are always due to busy traffic. It is expected that including other sources such as construction noises and other transport-related contributions in such environments would lower the observed correlation. Additionally, the diversity of available isolated source extracts for scene simulation is rather low. Particularly, voice events are recordings of read english with low variations in speaker and consistently neutral expressiveness. These properties result in no significant correlation of indicators describing voice events to pleasantness for this corpus.

The $\hat{P}_{1,\varphi}$ model is compared to two baselines proposed in [24] and [11], noted $\hat{P}_{2,\varphi}$ and $\hat{P}_{3,\varphi}$ respectively, for which predictor variables are directly computed from the audio signal. Coefficients for both models are re-optimized on the studied data for a fair comparison.

$$\hat{P}_{2,\varphi} = 18.67 - 0.20\ L_{50} - 0.02\ (L_{10} - L_{90}) \tag{7}$$

$$\hat{P}_{3,\varphi} = 30.18 - 0.16\ L_{50,1kHz} + 8.92\ TFSD_{500Hz,1s} + 2.99\ TFSD_{4kHz,1/8s} \tag{8}$$

The $\hat{P}_{2,\varphi}$ model does not explicitely involve the contribution to pleasantness of specific sources, but rather of emerging sound events with the $L_{10} - L_{90}$ indicator. Conversely, the $\hat{P}_{3,\varphi}$ model includes the time and frequency second derivative in the $500Hz$ and $4kHz$ bands at relevant time scales to underline the activity of voice and bird sources.

Table 5 summarizes the performance of the perceptual and physical models. The perceptual model $\hat{P}_{1,p}$ yields a root mean squared error of 0.61, which is below the average standard deviation of pleasantness assessments for this experiment: 1.77 on a 11-point scale. $\hat{P}_{1,\varphi}$ outperforms both $\hat{P}_{2,\varphi}$ and $\hat{P}_{3,\varphi}$, though a validation on a different corpus would be necessary to conclude on its capabilities. Compared to the model from perceptual parameters its R^2_{adj} is about 9% lower. Its root mean squared error is 0.83, which is also high compared to the perceptual baseline, but acceptable considering the average standard deviation of pleasantness assessments of 1.77 in this study. Introducing the $L_{10} - L_{90}$ emergence in $\hat{P}_{2,\varphi}$ has no impact for the considered corpus, as the same overall performance metrics were observed using only the L_{50}.

Table 5: Performance of baseline models for pleasantness prediction (**: p<0.01).

	$RMSE$	R^2_{adj}	r
$\hat{P}_{1,p}$	0.61	0.90	0.95**
$\hat{P}_{1,\varphi}$	0.83	0.82	0.91**
$\hat{P}_{2,\varphi}$	0.90	0.79	0.89**
$\hat{P}_{3,\varphi}$	0.91	0.78	0.89**

4.2 Prediction using deep learning

The proposed deep learning method is applied to the perceptual experiment corpus to obtain estimations for the time of presence of traffic, voices and birds. These estimations are then applied to pleasantness estimation and compared to models presented in Section 4.1. To evaluate the model's robustness to the increased polyphony and source complexity of scenes in real-life scenarios, the listening test corpus is split in three parts: 1) the 6 recorded scenes which contain additional sources not present in the datasets simulated for the optimization of the deep learning model, 2) the 19 replicated scenes that also include additional sources as annotated in [43], and 3) the

Table 6: Pleasantness prediction quality on the perceptual experiment corpus using the source detection model compared to labels. The corpus is split in three parts: the 6 recorded scenes (Rec.), the 19 replicated scenes (Rep.), and the 75 scenes with simulated scenarios (Sim.).

Model	$P_{1,\varphi}$ with model outputs				$P_{1,\varphi}$ with $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ labels		
Sub-corpus	All	Rec.	Rep.	Sim.	All	Rep.	Sim.
RMSE	0.84	1.09	0.68	0.85	0.83	0.72	0.86
r	0.91**	0.89**	0.93**	0.89**	0.91**	0.92**	0.89**
R^2_{adj}	0.82	0.73	0.82	0.80	0.82	0.79	0.79

75 simulated scenes that are obtained from the same simulation process as both the development and evaluation dataset.

Pleasantness predictions are obtained by substituting time of presence estimates computed from the presence detection architecture's outputs to the $\hat{P}_{1,\varphi}$ model presented in Section 4.1. Thus, only the outputs corresponding to the presence of birds are used. Table 6 presents the performance of pleasantness estimations using outputs from the deep learning architecture compared to those using ground truth $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ labels computed with eq. 2 on separated source-specific channels. First, pleasantness estimations are equally effective using the model's predictions or the ground truth labels, with about 0.84 RMSE and 82% of explained variance on the perceptual experiment corpus. This performance of the detection model is expected given the low errors on time of presence estimates in Table 3. Labels from the detection model result in lower errors on the first sub-corpus containing replicated scenes with sources not seen during training. This result may indicate that the detection model generalizes well to additional sources, though a larger sample size would be required to confirm this interpretation. For the simulated subcorpus of the experiments all performance metrics are comparable. Applying the detection model on the corpus of recorded scenes for which ground truth pleasantness is available results in a RMSE of 1.09 and a decrease in R^2_{adj}. This is expected as these scenes are the most distant from the training corpus in terms of sources and scenarios.

Pleasantness prediction for simulated scenes using the detection model is as effective as the best baseline model from acoustical indicators in table 5. This indicates that detection errors found in Table 3 do not impact pleasantness prediction on average. Since the labels are extracted from a masking model approximating the perceived time of presence, they can be considered as "weak" labels. Thus, the deep learning model's predictions are in some cases considered erroneous but correlate better to perceptual assessments than the corresponding ground truth labels.

5 Discussion

This study shows the potential of deep learning architectures in combination with corpora of simulated sound scenes for the perceptual characterization of sound environments. With the rich additional information about the source composition available for such corpora, new indicators are computed that outperform existing ones in their relation to the perceived time of presence of sources. Training a deep learning architecture on a large corpus of simulated scenes automatically annotated using these indicators then allows for the prediction of the time of presence in new recordings, where information about the contribution of each source is not available. The resulting predictions can be applied to the estimation of descriptors of soundscape perception such as the pleasantness.

Though, the performance of the trained model and its capacity to generalize to recorded data rely on several conditions. First, the perceptual characteristics of the simulated corpus should be comparable to that of typical urban environments. According to the results of the listening test presented in Section 2.3, the simulation process of sound scenes does not significantly affect the relations between abstract or content-related subjective descriptors, even with newly generated scenarios and a reduced source taxonomy. This is further demonstrated by the model of pleasantness found in Section 4.1 from subjective annotations of source activity in this experiment. The perceptual models on the listening test corpus are similar to those found in previous studies with both *in situ* questionnaires [11] and laboratory experiments using recordings [24]. Second, the quality of predictions is bounded by that of annotations. Here, the proposed indicator does not perfectly correspond with the perceived time of presence obtained during the listening test. Furthermore, the indicator is computed using two parameters optimized on available subjective data, though it was found to be quite robust by using a cross-validation scheme during its optimization. While the quality of predictions from the model are encouraging, testing on a larger corpus is required to fully assess its generalization capabilities.

The methodology proposed in this paper can be extended to include the identification of additional sound sources, though it requires the availablity of iso-

lated samples to simulate enough diverse sound scenes to train the detection architecture. This is because the scene simulation process, the time of presence indicator, and the deep learning model are all independent from the considered taxonomy. Of course, the complexity of the learning process scales with the number of considered simultaneously active sources. Particularly, differentiating between sources with similar spectral shapes may yield higher error rates as the deep learning architecture relies on the identification of patterns in third octave spectra. To obtain sufficient amounts of data, data augmentation techniques can be used to diversify the training dataset by applying slight modifications to existing examples, such as filtering, pitch shifting or time stretching [35, 54].

Future work will thus focus on studying the robustness of the proposed prediction scheme to a refined sound sources taxonomy as well as its application to a large scale sensor network.

Acknowledgements

This research is funded by the French National Agency for Research, under the CENSE project (convention ANR-16-CE22-0012). Part of the data used was collected in the framework of the GRAFIC project, supported by the French Environment and Energy Management Agency (ADEME) under contract No. 1317C0028.

References

[1] J. Ardouin, L. Charpentier, M. Lagrange, F. Gontier, N. Fortin, J. Picaut, D. Ecotière, G. Guillaume, and C. Mietlicky. An innovative low-cost sensor for urban sound monitoring. In *INTER-NOISE 2018*, 2018.

[2] C. Mydlarz, J. Salamon, and J.P. Bello. The implementation of low-cost urban acoustic monitoring devices. *Applied Acoustics*, 117:207–218, 2017.

[3] ISO 12913-1:2014. Acoustics - soundscape - part 1: definition and conceptual framework. Standard, International Organization for Standardization, Geneva, CH, 2014.

[4] A. Can, L. Leclerq, J. Lelong, and J. Defrance. Capturing urban traffic noise dynamics through relevant descriptors. *Applied Acoustics*, 69:1270–1280, 2008.

[5] L. Brocolini, C. Lavandier, M. Quoy, and C. Ribeiro. Measurement of acoustic environments for urban soundscapes: choice of homogeneous periods, optimization of durations, and selection of indicators. *J. Ac. Soc. Am.*, 134:813–821, 2013.

[6] M.E. Nilsson and B. Berglund. Soundscape quality in suburban green areas and city parks. *Acta Acust. unit. Acust.*, 92:903–911, 2006.

[7] P. Lundén, O. Axelsson, and M. Hurtig. On urban soundscape mapping: A computer can predict the outcome of soundscape assessments. In *INTER-NOISE 2016*, 2016.

[8] J.-J. Aucouturier, B. Defreville, and F. Pachet. The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. *J. Ac. Soc. Am.*, 122:881–91, 2007.

[9] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen. Polyphonic sound event detection using multi label deep neural networks. In *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015.

[10] F. Gontier, M. Lagrange, P. Aumond, A. Can, and C. Lavandier. An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach. *Sensors*, 17, 2017.

[11] P. Aumond, A. Can, B. De Coensel, D. Botteldooren, C. Ribeiro, and C. Lavandier. Modeling soundscape pleasantness using perceptive assessments and acoustic measurements along paths in urban context. *Acta Acust. unit. Acust.*, 103:430–443, 2017.

[12] A. J. Torija, D. P. Ruiz, and A. F. Ramos-Ridao. Application of a methodology for categorizing and differentiating urban soundscapes using acoustical descriptors and semantic-differential attributes. *J. Ac. Soc. Am.*, 134:791–802, 2013.

[13] M.E. Nilsson, D. Botteldooren, and B. De Coensel. Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas. In *19th International Congress on Acoustics*, 2007.

[14] S. Viollon and C. Lavandier. Multidimensional assessment of the acoustic quality of urban environments. In *INTER-NOISE 2000*, 2000.

[15] O. Axelsson, M.E. Nilsson, and B. Berglund. A principal components model of soundscape perception. *J. Ac. Soc. Am.*, 128:2836, 2010.

[16] R. Cain, P. Jennings, and J. Poxon. The development and application of the emotional dimensions of a soundscape. *Applied Acoustics*, 74:232–239, 2013.

[17] J.Y. Jeon, J.Y. Hong, C. Lavandier, J. Lafon, O. Axelsson, and M. Hurtig. A cross-national comparison in assessment of urban park soundscapes in France, Korea, and Sweden

through laboratory experiments. *Applied Acoustics*, 133:107–117, 2018.

[18] F. Aletta, J. Kang, and O. Axelsson. Soundscape descriptors and a conceptual framework for developing predictive soundscape models. *Landsc. Urban Plan.*, 149:65–74, 2016.

[19] P. Delaitre, C. Lavandier, C. Ribeiro, M. Quoy, E. D'Hondt, E. Gonzalez Boix, and K. Kambona. Influence of loudness of noise events on perceived sound quality in urban context. In *Inter Noise*, 2014.

[20] C. Guastavino. Categorization of environmental sounds. *Can. J. Exp. Psychol.*, 61:54–63, 2007.

[21] B. Gygi, G. R. Kidd, and C. S. Watson. Similarity and categorization of environmental sounds. *Perception & Psychophysics*, 69:839–55, 2007.

[22] A.L. Brown. Towards standardization in soundscape preference assessment. *Applied Acoustics*, 72:387–392, 2011.

[23] C. Lavandier and B. Defreville. The contribution of sound source characteristics in the assessment of urban soundscapes. *Acta Acust. unit. Acust.*, 92:912–921, 2006.

[24] P. Ricciardi, P. Delaitre, C. Lavandier, F. Torchia, and P. Aumond. Sound quality indicators for urban places in paris cross-validated by Milan data. *J. Ac. Soc. Am.*, 138:2337–2348, 2014.

[25] J.Y. Hong and J.Y. Jeon. Relationship between spatiotemporal variability of soundscape and urban morphology in a multifunctional urban area: A case study in Seoul, Korea. *Building and Environment*, 126:382–95, 2017.

[26] C. Lavandier, P. Aumond, S. Gomez, and C. Dominguès. Urban soundscape maps modelles with geo-referenced data. *Noise Mapping*, 3:278–94, 2016.

[27] A. Mesaros, T. Heitolla, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. DCASE 2017 challenge setup: tasks, datasets and baseline system. In *Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 workshop*, 2017.

[28] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters*, 65:22–8, 2015.

[29] B. McFee, J. Salamon, and J. P. Bello. Adaptive pooling operators for weakly labeled sound event detection. *IEEE Transactions on Audio, Speech and Language Processing*, 26:2180–93, 2018.

[30] S. Adavanne, P. Pertila, and T. Virtanen. Sound event detection using spatial features and convolutional recurrent neural network. In *2017 IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2017.

[31] M. Boes, K. Filipan, B. De Coensel, and D. Botteldooren. Machine listening for park soundscape quality assessment. *Acta Acust. unit. Acust.*, 104:121–30, 2018.

[32] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley. Acoustic scene classification. *IEEE Signal Processing Magazine*, 32:16–34, 2015.

[33] K. Sun, B. De Coensel, K. Filipan, F. Aletta, T. Van Renterghem, T. De Pessemier, W. Joseph, and D. Botteldooren. Classification of soundscapes of urban public open spaces. *Landscape and Urban Planning*, 189:139–55, 2019.

[34] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen. DCASE 2016 acoustic scene classification using convolutional neural networks. In *Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 workshop*, 2016.

[35] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24, 2017.

[36] L. Yu and J. Kang. Modeling subjective evaluation of soundscape quality in urban open spaces: An artificial neural network approach. *J. Ac. Soc. Am.*, 126:1163–74, 2009.

[37] A. Mesaros, A. Diment, B. Elizalde, T. Heitolla, E. Vincent, B. Raj, and T. Virtanen. Sound event detection in the DCASE 2017 challenge. *IEEE Transactions on Audio, Speech and Language Processing*, 27:992–1006, 2019.

[38] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia*, 2014.

[39] J. F. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. Moore, M. Plakal, and M. Ritter. Audio Set: an ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2017.

[40] Grégoire Lafay, Mathieu Lagrange, Mathias Rossignol, Emmanouil Benetos, and Axel Roebel. A morphological model for simulating acoustic scenes and its application to sound event

detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(10):1854–1864, 2016.

[41] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello. Scaper: A library for soundscape synthesis and augmentation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[43] J.R. Gloaguen, A. Can, M. Lagrange, and J.F. Petiot. Creation of a corpus of realistic urban sound scenes with controlled acoustic properties. In *Meetings on Acoustics*, 2017.

[44] F. Gontier, P. Aumond, M. Lagrange, C. Lavandier, and J.-F. Petiot. Towards perceptual soundscape characterization using event detection algorithms. In *Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 workshop*, 2018.

[45] P. Dagnelie. *Principes d'expérimentation: planification des expériences et analyse de leurs résultats*. Presses Agronomiques de Gembloux, 2003.

[46] H. Møller. Fundamentals of binaural technology. *Applied Acoustics*, 36:171–218, 1992.

[47] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–3, 1945.

[48] J. W. Pratt. Remarks on zeros and ties in the Wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54:655–67, 1959.

[49] M. Berzborn, R. Bomhardt, J. Klein, J.G. Richter, and M. Vorlander. The ita-toolbox: an open source matlab toolbox for acoustic measurements and signal processing. In *43th Annual German Congress on Acoustics*, 2017.

[50] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *2017 Conference Neural Information Processing Systems (NIPS)*, 2017.

[51] D. P. Kingma and J. Lei Ba. Adam: a method for stochastic optimization. In *2015 International Conference on Learning Representations (ICLR)*, 2015.

[52] J. Blauert and U. Jekosch. Sound quality evaluation: a multi layered problem. *Acta Acust. unit. Acust.*, 83:747–53, 1997.

[53] U. Jekosch. Basic concepts and terms quality, reconsidered in the context of product-sound quality. *Acta Acust. unit. Acust.*, 90:999–1006, 2004.

[54] M. Lasseck. Acoustic bird detection with deep convolutional neural networks. In *Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 workshop*, 2018.