Analysis and synthesis of urban sound scenes using deep learning techniques

Félix Gontier

THESE DE DOCTORAT DE

L'ÉCOLE CENTRALE DE NANTES

ECOLE DOCTORALE N° 602
Sciences pour l'Ingénieur
Spécialité : Acoustique

Par
Félix GONTIER

Analyse et synthèse de scènes sonores urbaines par approches d'apprentissage profond

Thèse présentée et soutenue à Nantes, le 15 décembre 2020
Unité de recherche : UMR6004, Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Dick Botteldooren	Professeur, Ghent University, Belgique
Gaël Richard	Professeur, Télécom Paris, Palaiseau

Composition du Jury :

Président :	Catherine Marquis-Favre	Directrice de recherche, ENTPE, Vaulx-en-Velin
Examinateurs :	Romain Serizel	Maître de conférences, Université de Lorraine, Vandœuvre-lès-Nancy
Dir. de thèse :	Jean-François Petiot	Professeur des universités, École Centrale de Nantes
Co-dir. de thèse :	Catherine Lavandier	Professeure des universités, Université de Cergy-Pontoise
Encadrant :	Mathieu Lagrange	Chargé de recherche, École Centrale de Nantes

Acknowledgements

First and foremost, I would like to thank the members of my thesis committee, Catherine Marquis-Favre, examiner and jury president, Dick Botteldooren and Gael Richard, reviewers, and Romain Serizel, examiner, for the interest they demonstrated towards my work and their insightful comments.

I am deeply grateful to my thesis director Jean-FranÃ§ois Petiot, as well as Catherine Lavandier and Mathieu Lagrange who supervised my work. Each of them brought complementary expertise which was instrumental to the completion of my thesis, along with exceptional availability and helpfulness. In particular, Mathieu Lagrange guided my work with unparalleled open-mindedness, enthusiasm, and patience.

For providing me with a rich research environment, I am also grateful to the Sims team at Laboratoire des Sciences du Numérique de Nantes, as well as members of the Cense project. The diverse backgrounds and area of interest of the people I encountered in the past three years opened my eyes to numerous new perspectives that were important to my development as a researcher. Among them, special thanks to Pierre Aumond for his invaluable assistance on acoustic and psychoacoustic scientific fields.

I express my utmost appreciation to Vincent Lostanlen, who took the time to provide in-depth commentary on the present manuscript with considerable expertise.

I would like to thank Ranim Tom for his help in dataset curation, as well as the many students at École Centrale de Nantes who participated in listening experiments.

Lastly, my thoughts go to my family and friends, whose continued support largely contributed to making this doctoral study a pleasant experience.

Contents

List of Figures

List of Tables

List of publications

Publications dans des revues d'audience internationale à comité de lecture

- 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

Communications à des congrès internationaux à comité de sélection et actes publiés

- Gontier, F., Lagrange, M., Lavandier, C., and Petiot, J.-F. (2020). Privacy aware acoustic scene synthesis using deep spectral feature inversion. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

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

Introduction

Following the ever-increasing expansion of large cities observed in the last decades, almost 75% of European citizens currently live in urban areas. Despite being an important component of the quality of life, the sound environment is barely taken into account in the design of those areas. This has led to an important increase of noise pollution [2], which is responsible for several public health issues [3, 4, 5]. In order to mitigate current hazards and to be able to propose new ways that would incorporate the sound modality in the urban planning process, there is an important need to rigorously monitor and gather information about the acoustic environment in urban areas.

Subsequently, monitoring sound environments has become an active field of research, with studies mainly investigating noise assessment and reduction applications. In particular, such is the objective of the 2002/49/CE directive, which requires large European cities to maintain publicly available noise maps [6]. Beyond noise reduction applications, we believe that a development is needed in order to better characterize and control urban sound environments, by modeling and predicting their impact on the quality of life of urban residents and passers-by. To do so, it is necessary to measure sound environments, and infer meaningful information to communicate to citizens and influence the design of urban areas.

In terms of measuring sound environments, the recent advent of the Internet of Things (IoT) has enabled the development of large-scale acoustic sensor networks [7, 8]. Several projects have implemented such networks to monitor urban sound environments, where indicators describing the acoustic content are continuously measured and stored on servers to be processed into relevant quantities. Many of these projects are developed in the context of noise assessment studies, with improving predictive noise maps as the primary objective. For example, in the RUMEUR network, 45 sensors continuously gather acoustic information in Paris [9]. In addition, several short-term measurement campaigns specifically refine assessments of aircraft noise. The project also includes a study on graphical representations and car-

tography of sound environments to better communicate information about their quality to both citizens and administrative parties. The DYNAMAP European project [10] further investigates the potential of low-cost sensor networks in monitoring sound environments of large cities. The developed approach implements a limited number of sensors at important locations of road infrastructures throughout the target cities of Rome and Milan. The network aims at capturing the diversity of traffic conditions in the city, including traffic density, road type, and weather conditions. The gathered data corrects estimations of predictive sound propagation models to produce maps of short-term noise level variations [11]. The Cense project, to which the present work is applied, proposes a different approach where dense sensor grids are implemented in the main districts and streets of Lorient [12]. The sensor network comprises a larger number of low-cost sensors compared to previous studies. Part of the project focuses on data assimilation between predictive models and the spatially dense sensor measurements, with a study on the uncertainties of emission-propagation models. In addition to traditional noise maps, continuously recording informative acoustic data enables the study of more comprehensive approaches to the characterization of sound environments. In order to produce multimodal noise maps easily interpretable by the citizens, part of the project is thus dedicated to assessing the perceptual quality of measured sound environments by both residents and passers-by. To this aim, the soundscape approach is prevalent in recent studies on the perception of sound environments [13]. In this approach, soundscape quality can be modeled efficiently from the perceived activity of sources on short time periods. Continuous measurements in dense sensor networks should allow the prediction of these quantities of interest, in order to propose detailed short-term maps of sound quality.

Relatedly, recent developments in signal processing have resulted in new efficient tools to model audio signals. Deep learning methods now achieve state-of-the-art results in sound event detection and sound scene classification. The growing Detection and Classification of Acoustic Sounds and Events (DCASE) community specifically investigates this range of applications, and proposes a recurring challenge on several related tasks[1]. Some studies have also successfully applied deep learning approaches to monitoring sound environments in the context of large-scale acoustic sensor networks. For example, the SONYC project [14] implements a low-cost sensor network in New York City and develops automatic methods for urban sound tagging [15]. Large collected measurement datasets also enable techniques such

[1] http://dcase.community

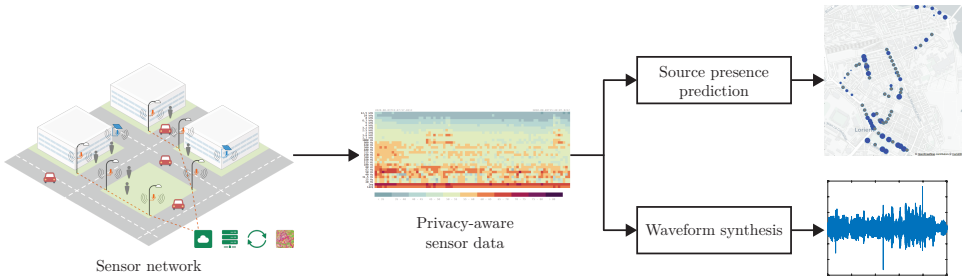


Figure 1: Overview of research topics where deep learning approaches are investigated in the current work.

as unsupervised representation learning for environmental audio [16].

This dissertation focuses on two contributions, that build on recently proposed deep learning techniques to i) predict perceptual attributes and ii) synthesize acoustic scenes from sensor data, as shown in Figure 1. We believe that the scientific advances in these two domains will allow to better inform citizens and city administrators about sound environments, by communicating easily interpretable information inferred from acoustic sensor networks.

Deep learning approaches to infer perceptual attributes of environmental sounds, either describing high-level properties (e.g. pleasant, lively) or the activity of sound sources, remain largely unexplored. The first contribution of this thesis is thus to propose an original approach using deep learning techniques to predict descriptors of perceived sound quality from sensor data. This includes the design of sufficiently large datasets and their annotation in terms of relevant underlying quantities describing the content of sound environments. In this context, recent studies in psychoacoustics identify the perceived activity of sources of interest as relevant descriptors.

Chapter 1 first establishes a summary of previous studies on the perception of environmental sounds and soundscape quality. After reviewing acoustic indicators that can be considered to represent perceptual quantities in published models, the potential of deep learning architectures in the estimation of high-level perceptual attributes is discussed. Specifically, the proposed original approach introduces predictions of perceived source presence from such architectures in models developed on large-scale studies on urban soundscape perception.

The information content recorded by sensors is subject to privacy regulations. Some applications enforce this constraint by obfuscating speech in voice activity segments obtained with source separation algorithms [17].

Instead, the Cense network anonymizes recorded data by directly measuring spectral sound levels [18]. The information content of measurements is insufficient to recover waveform audio with fully identifiable sound sources, and thus to manually annotate extracts for the desired task. Chapter 2 discusses the interest of sound scene simulation tools for automatic annotation of large datasets. The developed methodology allows the simulation of diverse sound scenes by combining extracts of isolated source occurrences according to realistic scenarios extrapolated from a corpus of annotated recordings. A listening experiment compares perceptual assessments of simulated scenes to results from the literature. An indicator is then proposed that models the saliency of a given source with respect to all others, from separated source contributions available in simulated scenes. This indicator is associated to the perceived presence of sources of interest in case of active listening of the sound environment by passers-by. Evaluation of the proposed annotation is done against state-of-the-art acoustic indicators, with respect to subjective assessments obtained in the listening test.

In Chapter 3, the developed indicator automatically annotates a large dataset of simulated scenes on the perceived presence of sources of interest. Two trained models, including convolutional and recurrent architectures, are able to predict source presence in synthetic data with high accuracy. Nonetheless, experiments evidence the incompleteness of simulated scenes with respect to sound environments of the application. In particular, the taxonomy of sound sources and associated spectral patterns in simulation processes may limit the adaptability of models to sound sources specific to target environments.

Chapter 4 addresses those limitations by localizing deep learning models to sound environments measured by the Cense sensor network. To do so, transfer learning approaches infer relevant information from large amounts of unlabeled sensor data in pretext tasks. Informative latent audio representations learned with these methods enable the training of low-complexity architectures predicting the perceived source presence. For this purpose, a second corpus of simulated sound scenes is constructed from on-site recordings of isolated source occurrences. The proposed experiments compare the relevance of latent audio representations obtained with two discriminative tasks relying on sensor metadata as well as an unsupervised regressive task. Evaluations done on a corpus of annotated recordings gathered in the target sound environments show that localizing simulated data highly contributes to prediction accuracy. Transfer learning and self-supervision further improves the effectiveness of architectures trained on simulated sound corpora with limited domain correspondence, data quantity, or diversity. The pro-

posed approach of perceived source presence prediction is adapted to the target *in vivo* sound environments, although future work remains regarding its potential embedding into low-cost sensors and its integration to the Geographic Information System (GIS) of the sensor network.

The second contribution aims at producing artificial sound environments that are conditioned with spectral information issued from the sensor network. Listening to waveform audio examples associated to sensor measurements is useful to provide citizens and urban planners with reference sound scenes illustrating noise maps. In some sensor networks, for instance the SONYC project, waveform audio is directly recorded and stored on secure servers. However, this is not always possible nor desirable due to regulations on privacy constraints. In the Cense project, sensor measurements sufficiently degrade the information content so that no waveform audio with intelligible speech can be retrieved [18]. This is also detrimental to the identification of other sound sources in reconstructed audio. In recent literature, deep learning architectures are successfully applied to the synthesis of speech (vocoders) and musical (musical synthesizers) signals. However, few approaches are applied to environmental sound synthesis, which contain complex polyphonies as well as diverse spectral signatures associated to sound source contributions. Thus, the second part of this thesis focuses on proposing deep learning models to synthesize plausible sound scenes that correspond to privacy-aware sensor measurements. Sound synthesis being a difficult task, the present work aims at proposing a preliminary study on original approaches to solve issues specific to environmental scenes processing. The developed methods thus require significant future work in order to achieve production-ready sound scene synthesis from sensor network measurements.

The use of deep generative models to synthesize plausible waveform audio from privacy-aware sensor data is investigated in Chapter 5. Two spectral approaches are considered where a magnitude spectrogram is reconstructed from measured third-octave sound levels, and phase information is recovered using iterative algorithms to obtain a waveform signal. Specifically, the first proposed architecture deterministically refines estimations from a data-independent baseline by learning *a priori* information from a dataset of acoustic scenes. The second proposed architecture upsamples log-frequency spectral representations, and is trained in an adversarial setting to enforce both the realism and the fidelity of generated spectra. The models are evaluated against deep and non-deep baselines available in the literature using spectral and waveform reconstruction metrics, as well as objective perceptual metrics and automatic classification performance.

We believe that those technical contributions will allow better citizen information and soundscape design in urban areas. The prediction of perceptual source activity descriptors is satisfying for the target environment, even if the proposed design and evaluation paradigms should be replicated in other urban areas. The availability of privacy-aware measurements through dense acoustic sensor networks allows exposing useful information about urban sound environments, and deep learning approaches are a promising avenue of research to do so.

The task of sound scene synthesis being inherently difficult, many underlying problems are still open in the deep learning community. The contribution developed in this thesis focuses on architectures enforcing invariance of regression loss functions to important variations in the magnitude of environmental sound spectra across frequencies. While experiments demonstrate the usefulness of the proposed approach, much work is still required to fully understand and tackle the fascinating problem of high-rate audio synthesis with deep learning paradigms.

Chapter 1

Perception of urban sound environments

The soundscape approach is prevalent in recent studies to qualify and quantify the perception of sound environments. This chapter reviews recent literature on the subject, including the characterization of soundscape quality by standardized perceptual dimensions and their relation to the activity of sound sources. In the context of continuous acoustic monitoring with sensor networks, these perceptual quantities can be efficiently approximated by acoustic indicators.
A framework is then proposed where such indicators are replaced by deep learning architectures. Specifically, the task of estimating the perceived time of presence of sources is formulated as an tractable problem of binary presence prediction on small temporal frames, related to event detection and classification tasks thoroughly investigated in deep learning communities. Predictions are introduced in linear models, validated in large-scale perceptual studies, to obtain estimations of high-level attributes of soundscape quality.

1.1 Soundscape quality

The impact of environmental sounds on society is traditionally associated with the notion of noise. This approach only addresses the negative implications of sound energy in an environmental setting. Thus, the acoustic quality of sound environments is inherently tied to the notion of annoyance. Particularly, in urban environments noise assessment is often reduced to traffic activity as the main contributor to sound levels and an important cause of annoyance. The concept of soundscape, initially proposed by [19], fundamentally differs from the noise approach. The soundscape is defined as "the acoustic environment as perceived or experienced and/or understood by a person or people, in context" [20]. In other terms, it represents the overall human perception of sound environments. The soundscape approach is thus a more comprehensive alternative to noise annoyance, that considers the ensemble of sounds that occur simultaneously with potentially positive effects instead of focusing purely on energetic aspects of the resulting sound environments. Despite its complexity, several studies have attempted to qualify the soundscape and its quality through sets of high-level perceptual attributes, or perceptual dimensions.

In [21], 27 descriptors of soundscape perception are evaluated in both *in situ* and laboratory conditions. A principal components analysis reduces this set to three dimensions that explain most of the variance in quality assessments, and respectively correlated to affective impressions and preferences (*i.e.* pleasant, comfortable, stimulating), activity due to sound presence of human beings (e.g. bustling, marked by living creatures, noisy) and auditory expectations (e.g. unexpected). Ten years later, a larger study is conducted in [22] with 116 subjective attributes. Likewise, a principal components analysis yields three main dimensions of soundscape perception. In the first two dimensions, axes are associated with the pleasantness and the eventfulness or liveliness, and explain 50% and 18% of the assessments variance respectively. Another space is obtained by applying a 45° rotation on the principal components space, in which axes are associated with the interest and calmness of sound environments. Similar results are also obtained in [23] where the two main dimensions are the calmness and vibrancy. The four axes in [22], corresponding to eight high-level attributes arranged in differential scales, are further proposed as part of a soundscape quality evaluation protocol for Swedish urban environments in [24]. In [25] the authors assess the challenges of standardizing this procedure to diverse cultures and languages by comparing results of laboratory listening tests in France, Korea and Sweden. Although the correspondence of assessments is verified for most

perceptual descriptors, the study concludes on the importance of carefully choosing translations as to not alter the meaning of descriptors. This subject is still investigated in recent work, with a translation of standardized English questionnaires for soundscape quality assessment [26] proposed in [27].

The two-dimensional model of pleasantness and liveliness is prevalent in recent literature. Nonetheless, some studies propose other descriptors of soundscape quality such as appropriateness [28] or music-likeness [29], with a review available in [30]. Furthermore, soundscape quality is increasingly associated with the dimension of pleasantness by itself [31].

1.2 Content-based assessment of soundscape quality

One of the important considerations stemming from the soundscape approach compared to the assessment of noise annoyance is the disparity between the sound level in an acoustic environment and its perceived quality. Noise annoyance does not discriminate between sound sources composing the environment, thus all contributions to the overall sound level are considered to affect perception negatively. A large-scale study is conducted in [32] by comparing *in situ* questionnaires with equivalent sound level measurements and the composition of sound environments in terms of sound objects in 14 urban environments across Europe, with assessments gathered for several years and different seasonal contexts. The observed relation between the measured sound level and the subjective acoustic comfort in this study is weak, whereas the activity of pleasant sound sources explains discrepancies of comfort evaluations in environments with equally high sound level. A similar conclusion on the importance of source types in soundscape quality assessment is reached in [33] from questionnaires answered by residents of French cities. Spontaneous descriptions of soundscape quality link positive judgments to nature, birds and most human activity sound sources, and negative judgments to mechanical sources such as traffic, cars and construction works.

In [34], a model of the unpleasantness is first constructed on a corpus of 20 stimuli evaluated in a laboratory setting. Evaluations include the perceived loudness as well as the presence, proximity and prominence (*i.e. combined presence and proximity*) of sound sources. Multiple linear regressions show that, in addition to the perceived loudness, significant negative contributions of traffic event sources (buses, mopeds) and a positive contribution of children voices to soundscape quality. Similar models obtained for *in situ*

evaluations also show positive contributions of bird sources, but underline changes in the perception of sources for different environments such as parks or streets. This phenomenon is further investigated in [31], by analysing about 3400 evaluations in diverse environments in Paris. The clustering of locations results in 7 distinct classes, for which models of soundscape quality are constructed independently. The authors also propose general models which they argue may be preferable in mapping applications, as ambiance classification uncertainties could compensate the differences in soundscape quality prediction performance. The best general model with 52% explained variance includes visual amenity as a predictor. However, in real situations visual and acoustic quality are naturally correlated [35, 36]. Laboratory studies are then necessary to decorrelate these factors in order to study the influence of visual parameters on sound quality, which is found to be much lower at around 10%. Alternatively, a second model only focusing on sound achieves 34% explained variance, and shows contributions of the perceived overall loudness as well as the activity of traffic, voices and bird sound sources to soundscape quality. The activity of each source is evaluated by its time of presence, that is the ratio of perceived presence from "rarely present" to "frequently present". The overall loudness and the time of presence of traffic sources contributes negatively to soundscape quality, whereas voices and birds have a positive effect.

This high-level sound source taxonomy in conjunction with the perceived time of presence as a source activity descriptor seems sufficient to model the pleasantness dimension of soundscape quality. In [1], a perceptual model of pleasantness is obtained for subjective assessments during soundwalks with the same predictors and slightly different coefficients. A multilevel variance analysis of the model further shows that the combination of the overall loudness and the time of presence of the three source types explain 90% of the variance in assessments related to the change of sound environments. This is also a significant improvement compared to addressing only the overall loudness, which yields 65% explained variance.

In [37, 22], sound source categories are generalized even further, to technological (e.g. traffic, construction, ventilation), human, and nature (e.g. birds, water, wind) sources. Models of soundscape quality and pleasantness constructed in these studies take the source dominance as the activity descriptor. However, the definition of dominance and its evaluation method differs across studies. In [22], the dominance is binary (0 or 1) and in most cases only one source category is regarded as dominant. Conversely, in [37] the dominance is evaluated on a 5-point scale from "never heard" to "completely dominating". The same scale is selected in [25] with a different sound

source taxonomy including separate evaluation of water, bird and wind activity. In general, refining the taxonomy by decomposing general source types into subclasses (e.g. types of vehicles or voice expressiveness), or introducing new sources may be useful. For instance, the sounds of a water fountain can indirectly contribute to pleasantness by masking traffic noise to perception in some environments [38]. Other sources may also contribute to high-level subjective attributes besides the pleasantness, and should be investigated in future studies.

1.3 Acoustic indicators

The modeling of perceptual soundscape quality requires subjective inputs from *in situ* experiments (e.g. soundwalks) or laboratory listening tests with reproduced sound environments. This approach is limited by the spatial and temporal locality of sound environments investigated in the conducted experiments. Consequently, characterizing the perception of sound environments through objective descriptors is a growing interest of the community. In particular, the availability of low-cost acoustic sensor networks could enable continuous perceptual monitoring of the soundscape. In part motivated by the 2002/49/CE directive, several projects such as RUMEUR [9], DYNAMAP [10], SONYC [14] or CENSE [12] recently implemented sensor networks in large cities. Sensor data are typically composed of spectral energy measurements relevant to the development of cartography applications, as shown in Figure 1.1. Acoustic indicators could be derived from these measurements to construct predictive models of soundscape quality, with applications in evaluation and design of acoustical properties of urban spaces.

Current acoustic monitoring applications mainly focus on the negative impact of noise, quantified by indicators of sound energy. Maps produced in compliance with the 2002/49/CE directive [6] are based on the L_{den} indicator. The L_{den} is computed as the average of A-weighted sound levels during day, evening and night, weighted by the duration of each period (resp. 6h-18h, 18h-22h, 22h-6h). Evening and night sound levels are penalized by 5dBA and 10dBA respectively to represent the additional disturbance generated by noise during these periods. Although the L_{den} is useful in long-term noise-only approaches, it does not contain information about the variations in content associated with perceptual properties of the soundscape. Monitoring applications widely rely on other energy indicators such as the LA_{eq}. However, they also fail to describe the short-term temporal variations in the

sound environment, and are too sensitive to noise peaks when aggregated over shorter periods of time [39].

Other indicators have been introduced in studies that mostly focus on the description of traffic noise. First, derivations of energy indicators can quantify temporal variations in the signal, for example the moments of the short-term sound level, or statistical descriptors. Statistical descriptors mostly consist in percentile values of the sound level distributions over time, and provide information on the average event, background or overall levels, or emergence of sound events (resp. LA_{10}, LA_{90}, LA_{50}, $LA_{10}-LA_{90}$). Emergence-based descriptors also describe sound events, including the number of noise events and mask index, corresponding to the number of occurrences and cumulative time where the short term sound level is greater than a threshold value [40]. Some indicators are based on spectral contents of the sound. For instance, the $LC_{eq}-LA_{eq}$ provides information on low-frequency content [37] and the spectral center of gravity evaluates the overall spectral content [41]. Lastly, psychoacoustic indicators based on the perception of amplitude modulations (e.g. roughness, loudness) or spectral density (e.g. sharpness) are found to characterize well the subjective annoyance of traffic noise [42].

To find which set of acoustic indicators is useful for general soundscape characterization, several studies propose to classify soundscapes with clustering algorithms. The obtained classes are validated against subjective evaluations and an optimal set of indicators is extracted that contributes the most to classification. In [43] a set of statistical descriptors composed of percentiles values of sound levels and psychoacoustic indicators successfully

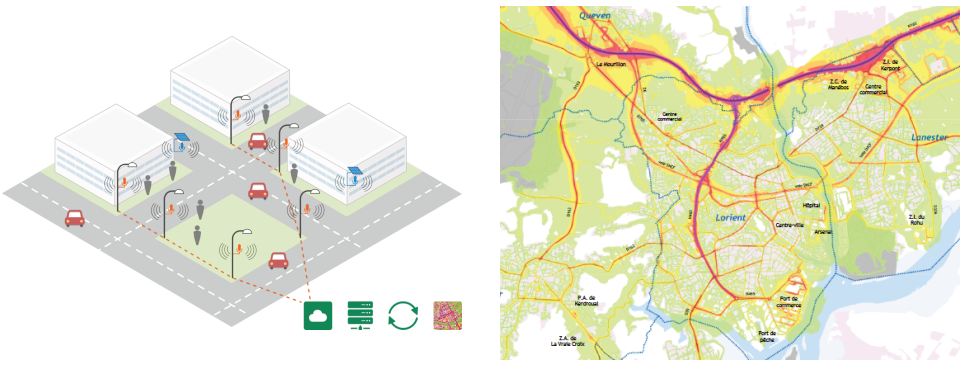


Figure 1.1: Example of sensor network developped as part of the CENSE project, and noise map (L_{den}) for the city of Lorient.

categorizes recordings of 370 soundwalks. In [44] 49 acoustic and psychoacoustic indicators are similarly investigated, and band-specific sound levels are found to outperform statistical descriptors in the optimal set. In [45], a limited set is sufficient to categorize sound environments, including the LA_{eq}, its standard deviation to describe temporal variations, and the spectral centroid describing spectral content.

In a soundscape quality prediction context, direct modeling of soundscape quality descriptors from sets of acoustic indices has also been studied. As shown in [46], energy-based indicators (L_{eq}, LA_{eq}) and psychoacoustic indicators (loudness, sharpness) correlate well with the unpleasant character of some sound environments, but not with pleasant evaluations of others. [37] identify energy-based indicators, specifically the LA_{50}, as strong predictors of the soundscape quality. However, the significant contribution of perceived nature and technological source activity is not well described by existing indicators for spectral content and temporal variations. Similarly, in [31] the best linear regression model of soundscape quality is obtained with the L_{50}, correlated to the perceived overall loudness, and the $LA_{10} - LA_{90}$. The explained variance of this model ($R^2_{adj} = 0.21$) is lower than that of the general perceptual model on the same corpus described in Section 1.2 ($R^2_{adj} = 0.34$), indicating the lack of acoustic indicators accurately describing perceived source activity. In [1] the time and frequency second derivative (TFSD) is introduced as a potential source-specific activity descriptor. The TFSD is computed as the discrete derivative, *i.e.* variation, of a third-octave spectrum $L(f, t)$:

$$TFSD_{f,t} = \frac{\left|\frac{d^2L}{df\,dt}\right|(f,t)}{\sum_{f_1=31.5Hz}^{f_1=16kHz}\left|\frac{d^2L}{df_1\,dt}\right|(f_1,t)} \tag{1.1}$$

where f and t denote the frequency and time dimensions of the spectrum. The scale of the time dimension can be changed by aggregating information on longer time scales before computing the TFSD. Thus, the TFSD can highlight spectral and temporal variations at different levels. For example, bird activity is associated with fast variations in high frequencies. The TFSD indicator computed at a 125 ms time scale and for the 4 kHz third-octave band is found to correlate well to the time of presence of birds evaluated during a soundwalk. To a lesser extent the TFSD computed with 1 s measurements for the 500 Hz band correlates with the perceived time of presence of voices. Both indicators are validated in a multiple linear regression model

of pleasantness with the $L_{50,1\,\mathrm{kHz}}$ as predictor for the perceived overall loudness and no specific predictor for traffic activity. TFSD variants outperform the compared reference indicators, including the number of sound events, the spectral center of gravity, and emergence descriptors.

Recent work also investigates the potential of machine learning methods in estimating perceptual responses to sound environments. In [47], neural networks are trained to directly predict soundscape quality from acoustic data in 19 urban environments. Networks trained for individual locations perform well but no general model is found. Support vector regression models are trained in [48] to predict the pleasantness and eventfulness of acoustic scenes. A bag-of-frames approach with Mel-frequency cepstral coefficients as input features yields estimations within the variance of individual participant assessments. The authors of [49] design a deep learning model to predict which sound sources are likely to be identified within an urban park environment. The architecture includes a recurrent component to model the perception of subsequent events. Linear regression models of perceived mechanical, nature, and human source activity, as well as soundscape quality, are then constructed with model predictions and other acoustic indicators as predictors.

Machine learning tools could also prove useful for isolating source-specific content. For instance, a source separation method was recently proposed in [50] for the estimation of traffic sound levels in urban environments. Source separation using deep learning models has also been successfully applied in speech and music domains, but remains relatively unexplored for environmental sounds.

State-of-the-art approaches converge towards the content-based modeling of perceptual quality. The current best performing indicators attempt to identify individual sound sources within the mix by filtering time-frequency representations of the audio signal. For example, the TFSD indicator underlines variations in a spectral representation at time scales representative of the sources of interest. However, this type of acoustic indicators is sensitive to sound sources overlapping in time and frequency, particularly sources characterized by similar time-frequency modulations but different perceptual implications. These issues could be mitigated with more complex models to identify content of interest within sound environments. Deep learning approaches provide the possibility of developing arbitrarily complex, highly nonlinear models able to capture the differences between patterns characterizing objects within a mix. With a well-motivated architecture, a deep learning model could thus efficiently identify sound sources based on time-frequency modulations, and link them to perceptual quantities without the

need for handcrafted acoustical descriptors.

1.4 Framework for soundscape quality prediction with deep learning

The present work aims at improving upon acoustic indicators presented in Section 1.3 with deep learning models, in order to predict the soundscape quality attributes in a large-scale sensor network context. Generally speaking, a deep neural network is a parametric transformation applied to an input x to predict an output y:

$$y = f_w(x) \tag{1.2}$$

Typical architectures are composed of an arbitrary number of linear transformations, either linear projections or filterbanks, each followed by the application of a nonlinear activation function to the output data.

$$\begin{aligned} h^{(1)} &= g^{(1)}(f_{w^{(1)}}(x)) \\ h^{(i)} &= g^{(i)}(f_{w^{(i)}}(h^{(i-1)})) \\ \hat{y} = h^{(N)} &= g^{(N)}(f_{w^{(N)}}(h^{(N-1)})) \end{aligned} \tag{1.3}$$

where $f_{w^{(i)}}$ is linear with learned parameters w_i, $g^{(i)}$ is a differentiable nonlinear function and N is the number of layers in the model. In a supervised learning setting, the prediction $\hat{y}$ is compared to a ground truth label y using a differentiable loss function f_L, which is minimized for optimal parameters w of the model:

$$L = f_L(\hat{y}, y) = f_L(f_w(x), y) \tag{1.4}$$

$$w_{opt} = \arg\min_w f_L(f_w(x), y) \tag{1.5}$$

As the transformation f_w is highly non linear, the loss is non-convex and direct optimization is difficult. However, the transformation is always differentiable. Thus, the parameters are optimized by stochastic gradient descent methods [51] for batches of examples of paired inputs and outputs $\{x_i, y_i\}_{i=1}^{M}$, where M is the number of available examples in the training dataset.

Because the architecture design of deep neural networks are arbitrary, they are regarded as universal approximators [52]. Despite being mainly applied to the extraction of physical information from input representations in the literature, the developed architectures are generally task-agnostic. Furthermore, they perform particularly well on discriminative tasks where a

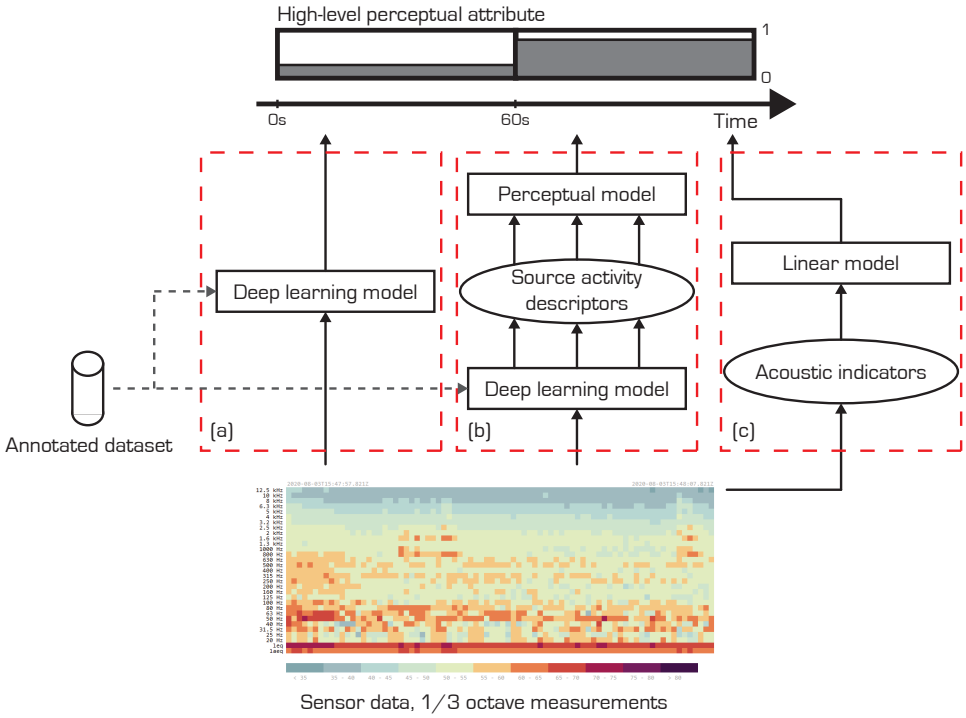


Figure 1.2: Possible configurations for the prediction of soundscape quality attributes from sensor data. (a) Direct prediction of high-level attributes. (b) Prediction of intermediary source activity descriptors and application of a perceptual model. (c) Approach developed in the literature, derivation of acoustic indicators and linear modeling of perceptual attributes.

low-dimensional output is predicted from a high-dimensional input representation. In the present study, a deep learning model can thus be introduced in a framework predicting perceptual soundscape quality attributes from sensor data, with possible configurations shown in Figure 1.2. A naive formulation of the soundscape quality prediction task consists in directly estimating high-level attributes from third-octave sound levels (Figure 1.2 (a)). In this setting, the model needs to both extract the relevant spectral patterns from the input representation and learn their impact on the desired output. This is a complex task as the model is given no prior information on which spectral patterns are relevant to modeling the attribute. On the other hand, the relationships between some soundscape quality attributes such as the pleasantness and descriptors of perceived source activity are available in the form of linear models discussed in Section 1.2. In the second setting shown in

Figure 1.2 (b), the deep learning model predicts intermediary source activity descriptors. These predictions are then fed to linear perceptual models to obtain estimates of high-level attributes. Although source activity descriptors are still perceptual in nature, their correspondences to time-frequency structures in the signal should be more straightforward to unveil. Thus, training a model to predict source-specific descriptors as intermediary dimensions of soundscape quality is preferable. This allows the introduction of predicted quantities in well-established perceptual models, developed in the literature with the configuration in Figure 1.2 (c), and also enables introducing more complex models or adaptations for specific environments in future research. Furthermore, source activity is useful as additional information easily interpretable by the citizen [53]. In this study, the dimension of pleasantness is specifically addressed, as it is increasingly associated to soundscape quality in recent studies. Perceptual models from source activity assessments are thoroughly investigated in the literature [31], providing a strong reference to evaluate the proposed approach. However, the proposed approach could be applied similarly to other high-level attributes.

The choice of a source activity descriptor then conditions the task that the deep learning model has to solve. In perceptual studies of Section 1.2, source activity is generally evaluated on discretized scales corresponding to continuous quantities. This is for example the case for the perceived loudness or emergence of sources of interest. The associated prediction task amounts to a statistical regression by nature, and typically requires large amounts of training data covering the range of possible input and output distributions. The estimation problem can be cast to a more tractable single-label classification task by quantizing the descriptor scale, and regarding each quantization step as an independent class. This technique has for example been applied to the prediction of individual audio samples in the WaveNet architecture [54]. However, classification loss functions such as the cross-entropy do not inherently account for class proximity, which can lead to additional difficulties in the training process [55]. The time of presence is also a quantitative and continuous scale, bounded between never heard (0) and always heard (1). However, its prediction may be formulated as a detection or classification task under the assumption that perception over time is stationary. That is, the perceptual time of presence can be obtained as the aggregation over time of the binary presence (absent-present) of sources of interest, evaluated at time scales relevant to perception. In this case, the problem is formulated as a multiple-label classification task where each class is associated to a sound source of interest, and multiple labels (*i.e.* sound sources) can be active simultaneously. This is akin to object recognition in image processing, for

which deep learning models are known to perform well.

Chapter conclusion

Recent studies model high-level perceptual attributes of soundscape quality from descriptors of source activity. We believe that deep learning approaches can efficiently infer these content-based descriptors from acoustic measurements. In particular, the perceived time of presence of sound sources is of interest. By approximating the time of presence as the aggregation over time of the perceived source presence in relevant temporal segments, its prediction amounts to a tractable multilabel classification task. Chapters 2 through 4 thus investigate an approach where deep learning models predict the perceived source presence from sensor measurements, and estimates of high-level attributes (e.g. pleasantness) are obtained with well-established perceptual models.

Chapter 2

Automatic annotation of perceived source activity for large controlled datasets

Training deep learning models to predict the perceived presence of sources of interest requires large datasets of labeled sound scenes. As an alternative to manually annotating recordings, controlled datasets of simulated acoustic scenes enabling automatic annotation are discussed.
A listening test is first conducted to verify that sound scenes simulated with the proposed method yield similar perceptual properties compared to recordings, both in terms of the relation between high-level attributes and their behavior with respect to content-based perceptual descriptors.
An emergence indicator is then developed to derive annotations of perceived source presence on short texture frames in simulated scenes. On the listening test corpus, this annotation correlates better to perceptual descriptors of traffic, voice and bird activity than state-of-the-art acoustic indicators proposed in the literature.

2.1 Introduction

Training a deep learning model to predict the perceived presence of sources as described in Section 1.4 requires annotated data in sufficient amounts. Supervised classification datasets are typically composed of several hours of audio [56]. For recordings of sound environments, labels of perceived presence for the sources of interest over time could be manually annotated by a small panel. However, this process is time-consuming and thus not scalable, as extending the scope of the study to include additional sources or environments requires repeating the annotation process for all new data. Alternatively, sound scenes can be automatically annotated with an acoustic indicator correlated with the perceived presence of sources. The predictions of trained deep learning models are at best of the same quality as that of labels in the training dataset. Thus, to justify automatic annotations, they should correlate better with the perceived time of presence of sources than state-of-the-art acoustic indicators in Section 1.3 such as the TFSD indicators [1]. As a result, training a deep learning model with presence labels extracted using acoustic indicators that can be computed on third-octave sensor data has no benefit compared to directly applying such indicators to the prediction of perceptual attributes. However, only the training process requires presence labels to optimize model parameters. An automatic method for annotating source presence can therefore be based on any information available on the training corpus, even if equivalent information is not available when applying the learned model to sensor data at evaluation.

In this context, sound scene simulation tools provide a great level of control over the composition of sound scenes by design, and thus ground truth knowledge about the physical activity of sources of interest. From this information, a robust and efficient perceived source presence indicator can be derived to automatically annotate a large training dataset.

Figure 2.1 shows the sound scene construction process for the two main simulation libraries available: *simScene* [57] and *Scaper* [58]. *Scaper* and the *replicate* mode of *simScene* function similarly: scenes are simulated based on an input scenario describing the background and event activity of sources of interest over time. Background sources are active for the duration of the sound scene. The first background (class *traffic* in Figure 2.1) constitutes the reference in terms of sound level. Additional background sources (class *crowd* in Figure 2.1) are associated with an event-to-background ratio (EBR) corresponding to the emergence of the event compared to the overall background activity (expressed in dB), and from which the corresponding extracts are scaled.

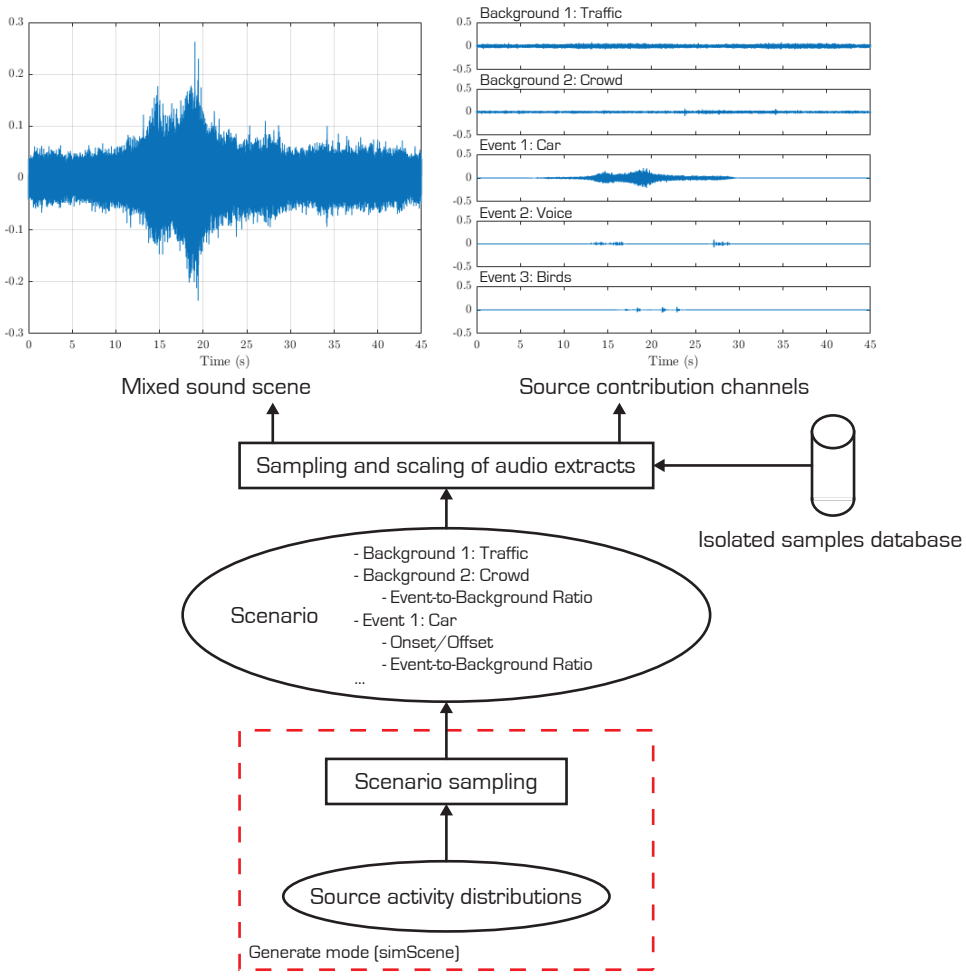


Figure 2.1: Overview of the scene simulation process from scenarios and a database of isolated source samples.

Event sources are defined by onset and offset timestamps, as well as an event-to-background ratio with all combined background sources as reference. From the input scenario, source contributions are pseudo-randomly sampled from a database of isolated samples, which comprises extracts of isolated occurences of a specific source. The simulation process outputs contributions of each source as separate channels, and combines them additively to produce the sound scene. For sound event detection (SED) datasets creation, *Scaper* additionally allows data augmentation on the isolated samples database with pitch shifting and time stretching techniques.

Figure 2.2: Map of the soundwalks and 19 recording locations in the 13th district of Paris presented in [1]. Sound levels shown on the soundwalk path are interpolated from measurements at each location.

The *generate* mode of *simScene* is oriented towards creative use cases, as it allows the generation of original scenarios. The term scenario refers to an ensemble of all properties characterizing sounds in polyphonic scenes, including the taxonomy of sources and structural parameters describing their activity [59]. Instead of scalar onset-offset and event-to-background ratio values, the simulation process in this mode is given the mean and standard deviation of a normal distribution for each source and simulation parameter. Original scenarios are composed by pseudo-randomly sampling these distributions. This process can be seen as performing data augmentation on a reference (average) sound scene scenario by applying random variations to its defining high-level properties, where these variations are controlled to remain plausible. Conditioning distributions on ambiances, *i.e.* inferring a different set of reference sound scenes as well as variation range for each category of sound environment, then results in a large dataset of sound scenes covering diverse scenarios encountered in urban environments.

The capacity of a deep learning model to generalize to possible urban sound environments depends on the quality of the training data. In other terms, the simulated scenes composing the training dataset should contain

diverse scenarios covering as many real-life situations as possible while remaining within the scope of plausible environments. To do so, the corpus of 74 recordings proposed in [1] is taken as a reference for the parameters of the scenario generation process.

This reference corpus, on which the TFSD indicator described in Section 1.3 is developed, was gathered as part of the GRAFIC project during four soundwalks in the 13th district of Paris. Sound scenes ranging from 55 s to 4.5 min in duration were recorded at 19 locations (P1-19) with diverse environments as shown in Figure 2.2. These recordings were classified in [60] as representing *quiet street*, *noisy street*, *very noisy street* and *park* ambiances.

2.2 Subjective annotations

2.2.1 Motivation

The capabilities of *simScene*'s *generate* mode are useful for creating large datasets with diverse ambiances and polyphonic levels, but raise some concerns about perceptual responses produced by simulated scenes. A deep learning model will be trained to predict perceptual descriptors on simulated scenes, then applied to real-life sensor data derived from recordings. There is thus a need to ensure that intrinsic differences between simulated and recorded scenes do not result in significant changes in terms of perception, both overall and in terms of active sound sources. Such differences exist at two levels: the additive composition of simulated scenes does not fully reflect the propagation and interactions of sound sources in real-life conditions, and the scenarios, although derived from existing environments, are original. The effect of the first factor on perception can be studied by manually annotating reference recordings, and using the resulting scenarios to simulate sound scenes with almost identical content in terms of source activity. Differences between perceptual evaluations of paired recorded-replicated sound scenes on individual descriptors can then be investigated. Assessing the impact of the second factor is less straightforward, because there is no direct correspondence between real-life and generated scenarios to compare quantitative assessments on. Evaluation should thus be conducted on a larger corpus to assess potential changes in the behavior of perceptual quantities.

Furthermore, the performance of an indicator proposed for the automatic annotation of perceived source presence in simulated scenes has to be evaluated with respect to subjective annotations of the time of presence. Lastly, the quality of high-level attributes estimated from predictions of a deep learning model as described in Section 1.4 should be evaluated in the

same manner.

Addressing these concerns thus requires perceptual annotations on a corpus of sound scenes. To this aim, a pilot study is first conducted with a listening test on a limited corpus of simulated scenes, that leads to the proposition of the perceptual time of presence annotation indicator presented in Section 2.4 [61]. This study is followed by a second test with the same design and objectives conducted on a larger simulated corpus [62], which is reported in this document.

2.2.2 Listening test corpus

To fulfill the objectives described in Section 2.2.1, a corpus of 100 sound scenes is constructed for the listening test. This corpus is composed of 6 recorded scenes, 19 replicated scenes and 75 simulated scenes. A duration of 45 s is chosen for all sound scenes to reduce participant fatigue during the listening test. Although many studies dealing with environmental acoustic quality rely on acoustic measurements ranging from a few seconds [63, 64] to 15 min [65, 66] and even to 80 min [67], stimuli between 30 s and 60 s are often preferred for laboratory tests.

First, the 74 scenes in the reference corpus are manually annotated in terms of traffic, human voice, and bird background and event activity. For sound events, annotations include the onset and offset as well as the event-to-background ratio (EBR). In the case of multiple sources active in the background, one is taken as reference and the event-to-background ratio for subsequent sources is also estimated. To do so, the recorded scenes are replicated with simScene's *replicate* mode from annotated event onsets-offsets and initial guesses of the event-to-background ratios as inputs. Each replicated scene is compared to the reference recording by informal listening and the EBR for each sound event is adjusted until they correspond.

The 19 replicated scenes corresponding to one of the four soundwalks in [1] are selected as part of the listening test corpus. 45 second segments are selected such that they do not contain one single overwhelming event. For 6 of the 19 replicated scenes (P1, P3, P4, P8, P15 and P18), the matching 45 second segments are extracted from the recordings and included in the listening test corpus for paired analysis. These 6 locations represent all of the 4 ambiances found in [60], with sound levels ranging from 63.9 dB SPL to 79.4 dB SPL.

Original scenarios of simulated scenes are then generated with simScene's *generate* mode. As described in Section 2.1, the scenario generation tool takes the first background source as reference, and samples an event-to-

background ratio applied to subsequent backgrounds from a normal distribution. For events of a given source, onsets are sampled successively from an input distribution of inter-onsets. Similarly, the event-to-background ratio is sampled independently for each event from a normal distribution. In this study the input distribution parameters as well as the probability of appearance are obtained for each source and ambiance by summarizing annotations of the 74 recordings in the reference corpus. The source taxonomy is limited to traffic, human voices, and birds as the main sources of interest in the estimation of pleasantness in Section 1.2, and because too few occurrences of other sources are present in the reference corpus to extract meaningful statistics. Each simulated sound scene thus contains at most six distinct source contributions, corresponding to either the background or event activity of one of the three sources in the taxonomy. The isolated samples database is constructed from a subset of the LibriSpeech [68] corpus for voice events and Freesound contributions for remaining sources. The isolated samples database contains 8 min, 17 min and 37 min of background traffic, voice and bird extracts respectively, as well as 3 min, 37 min, and 5 min of event traffic, voice and bird extracts respectively. No neutral background noise is added to simulated scenes, although several extracts in the isolated samples database contain uncontrolled noise components. During the simulation process, extracts corresponding to sound events are faded in and out for 10% of their duration to avoid unrealistic sudden cuts.

A fifth *square* ambiance is introduced with predominant voice activity. Scenario generation parameters for this ambiance are derived empirically from other ambiances as it is not represented in the reference corpus. For other ambiances, the standard deviations of event inter-onsets and event-to-background ratio distributions are increased to maximize the diversity of generated environments. As this may lead to implausible scenarios, for example with multiple events overlapping or overly loud birdsong, the realism of all simulated scenes is informally checked. Each simulated scene is further associated to a sound level sampled from a normal distribution conditioned on the ambiance. A summary of the parameters conditioning the scenario generation process is shown in Appendix A.

From these parameters, 200 sound scenes of 45 s each are first simulated with equal distribution over the five available ambiances (resp. *park*, *quiet street*, *noisy street*, *very noisy street* and *square*). 75 scenes are then selected as part of the listening test corpus. To do so, the indicator presented in Section 2.4.1 and initially proposed in [61] provides an estimate of the perceived time of presence of sources, between 0 (no presence) and 1 (presence 100% of the time). This indicator is computed for each of the 3 sources of interest,

and the resulting values are treated as coordinates in a 3-dimensional space. The 75 scenes that maximize the minimum pairwise Euclidean distances in this space, *i.e.* the 75 most isolated scenes, are selected. Figure 2.3 shows the distribution of selected simulated scenes. Note that not all ambiances are represented by the same number of simulated scenes in the final listening test corpus. Playback sound levels range from 46.6 dB SPL to 77.1 dB SPL over the 75 simulated scenes.

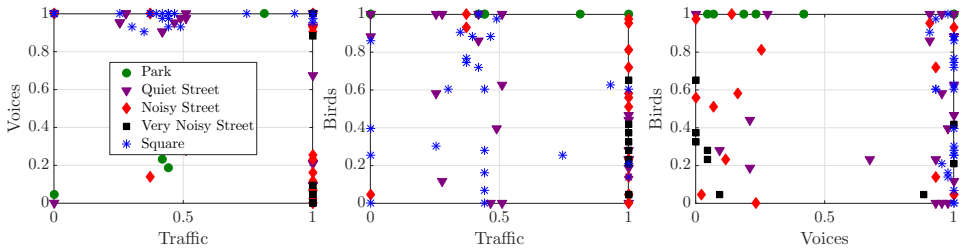


Figure 2.3: Time of presence of sources estimated by the indicator proposed in Section 2.4 for the 75 simulated scenes in the listening test corpus.

2.2.3 Subjective annotation procedure

Perceptual assessments are gathered on the corpus constructed in Section 2.2.2. During the listening test, participants are asked to evaluate each sound scene on 8 perceptual attributes. Each attribute is quantized on 11-point semantic differential scales (0-10). The first 4 descriptors are high-level attributes corresponding to the major and minor axes found in the first two dimensions of the principal components analysis of soundscape quality descriptors in [22]. They are translated using terminology from [25] and presented in French:

- Pleasantness: *Unpleasant - Pleasant* (*Désagréable - Agréable*),
- Liveliness: *Inert, amorphous - Lively, eventful* (*Inerte, amorphe - Animé, mouvementé*),
- Interest: *Boring, uninteresting - Stimulating, interesting* (*Ennuyeux, inintéressant - Stimulant, intéressant*),
- Calmness: *Agitated, chaotic - Calm, tranquil* (*Agité, chaotique - Calme, tranquille*).

The overall loudness is also evaluated as it appears in several perceptual models of soundscape quality:

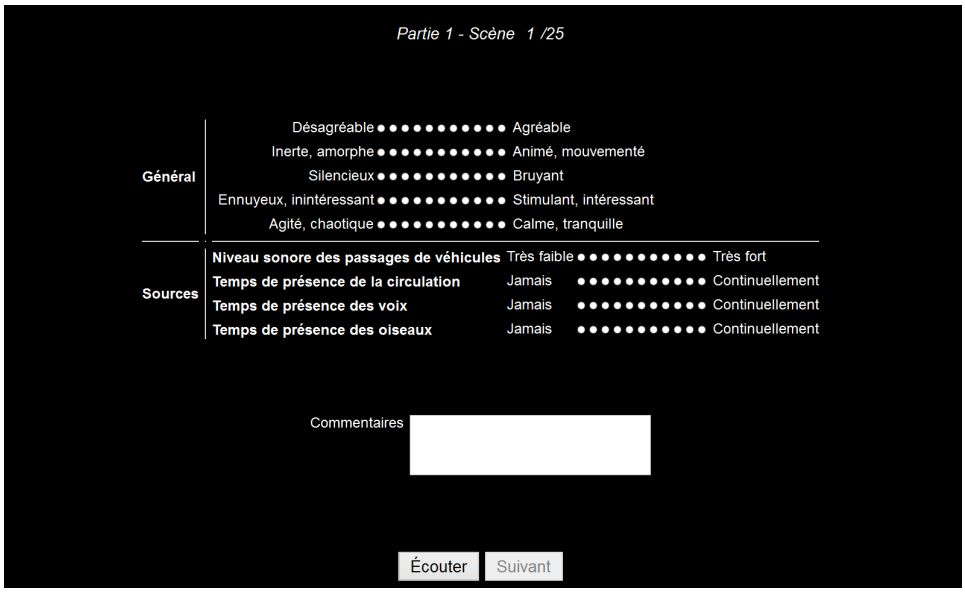


Figure 2.4: Screenshot of the Python interface presented during the listening test.

- Overall loudness: *Quiet - Noisy* (*Silencieux - Bruyant*).

Additionally, to assess the perceived source activity 3 questions are presented to the participants and evaluated on the same 11-point semantic differential scales:

- Time of presence of traffic, voice and bird sources: *Never - Continuously* (*Jamais - Continuellement*).

These time of presence assessments are referred to as $T_{T,p}$, $T_{V,p}$ and $T_{B,p}$ for traffic, human voice, and birds respectively in this study, where p denotes a perceptual evaluation. The listening test is conducted in a laboratory setting, and presented with a Python interface shown in Figure 2.4. The depicted sound level of passing vehicles scale is omitted in the analysis, as it is correlated to the perceived time of presence of traffic, and its potential is limited in the framework developed in Section 1.4. Furthermore, participants receive a short verbal introduction prior to the test to ensure that perceptual attributes are well understood.

A total of 23 students aged from 22 to 23 years (16 male students and 7 female students) at Ecole Centrale de Nantes completed the test, all reported normal hearing. All participants gave written consent prior to the

experiment, and evaluations were further anonymized. Each participant only listens to 50 of the 100 sound scenes to reduce hearing fatigue. They all evaluate the 6 recorded and 19 replicated scenes, and 25 of the 75 simulated scenes. The 25 scenes are chosen according to an incomplete block design [69], where each group of 3 consecutive participants cover all 75 simulated scenes. Thus, with 23 participants, the incomplete block design is not perfectly balanced: 50 of the 75 simulated scenes are evaluated by 8 participants whereas the remaining 25 are evaluated by 7 participants. Participants first evaluate the most quiet and loudest of the 6 recorded scenes, resp. P3 and P15 at 63.9 dB SPL and 79.4 dB SPL, to allow them to calibrate their subsequent ratings. The 48 remaining scenes are presented in random order. This allows us to control ordering effects over participants, *i.e.* a potential bias introduced by the hearing of previous scenes on the evaluation of the current scene. Furthermore, participants can only listen to an acoustic scene once, and must listen to the full extract and evaluate it on every attributes before proceeding to the next. However, they are allowed to start evaluating the scene before the end of the playback. The average duration of the test was about 45 min.

Each acoustic scene in the listening test corpus is associated with a sound level in dB at which the scene would be heard in real-life conditions. Thus, a calibration procedure is applied to ensure that sound scenes are heard at the desired sound level by every participants. All participants listen to sound scenes played through Beyerdynamics DT-990 Pro headphones. Furthermore, the scenes are played with the same computer sound card and software parameters, including Python libraries versions, sound card configuration, and software volume. The calibration of the headphones was carried out in a free field situation and consisted in characterizing the relationship between voltage at the headphone input and the corresponding binaural sound pressure. To do so, the following procedure was conducted in a semi-anechoic chamber:

1. A pink noise generator is set at an arbitrary level and its output RMS voltage V_{gen} is measured.

2. Small DPA 4060 binaural microphones are set at the entrance of the ear canals of a human participant [70]. The pink noise generator is input to the headphones, placed over the head of the participant. The RMS voltage at the output of the binaural microphones V_{bin} is measured.

3. The headphones are removed, and the generator input to a Genelec 1031A loudspeaker placed at a distance of 1 m from the participant's

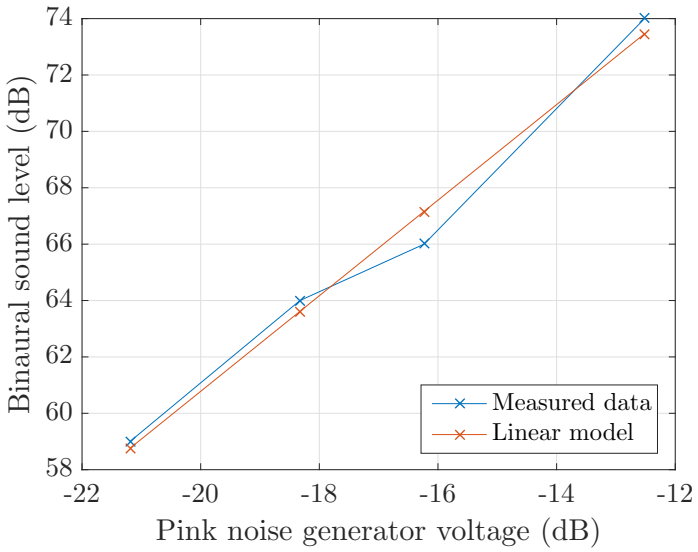


Figure 2.5: Measurements and linear model of the playback sound level of Beyerdynamics DT-990 Pro headphones (L_{head}) as a function of input pink noise electrical level ($log(V_{gen})$).

head. The loudspeaker's amplification is tuned until the same RMS voltage V_{bin} is measured at the output of the binaural microphones.

4. The head is replaced by a class 1 sound level meter measuring the sound level L_{head} (in dB) of the loudspeaker. This corresponds to the binaural sound level produced by the headphones for an output voltage of the pink noise generator V_{gen}.

By repeating this procedure for different settings of the generator level, the relation between the logarithm of the generator output voltage and headphones playback sound level in dB is obtained in Figure 2.5. This relation is approximated as a linear function:

$$L_{head} = 1.70 \cdot 20log_{10}\left(V_{gen}\right) + 94.7 \tag{2.1}$$

To complete the calibration for the listening test, a pink noise extract with RMS amplitude L_{num} (amplitude in the $[-1, 1]$ range) is played on the desired computer and the RMS voltage at the output of the sound card, equivalent to V_{gen}, is measured. Again, a linear relationship between V_{gen} and L_{num} is obtained for this hardware and software configuration. From

Table 2.1: Mean differences of perceptual assessments (resp. Pleasantness, Liveliness, Overall Loudness, Interest, Calmness, Time of presence of Traffic, Voices, and Birds) between recorded 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	$T_{T,p}$	$T_{V,p}$	$T_{B,p}$
P1	0.43	**-1.65**	**-1.04**	0.43	0.13	0.39	**-2.09**	0.61
P3	0.26	-0.43	0.30	-1	0.30	1.04	**-4**	0.22
P4	0.91	0	**-1.83**	0.48	**1.30**	**-5.22**	**1.43**	0.04
P8	0.26	**-1.65**	-0.87	-0.96	0.65	**-0.91**	0.09	**-1.43**
P15	**-1.35**	0.52	0.52	**-1.17**	0.09	0.13	**1.96**	**-2.74**
P18	**1.13**	-0.30	**-1.17**	-0.43	**1.39**	**-1.83**	**0.83**	**1.30**

these two models, a scaling factor is applied to each sound scene so that its RMS amplitude L_{num} results in the expected sound level L_{head}.

Subjective assessments obtained during the test are pre-processed by conducting an outlier detection procedure, where the mean and standard deviation of assessments for each of the 8 perceptual attributes and 100 sound scenes is computed. Participants with more than 10% of assessments differing from the mean by more than 3 standard deviations are removed from the study. This method does not evidence any outlier participant for this test.

2.3 Perceptual validation of simulated acoustic scenes

2.3.1 Simulated scene construction

The effect of the *simScene* simulation process by additive combination of isolated sound sources is first studied. To this aim, perceptual assessments for the 6 pairs of recordings and corresponding replicated scenes in the listening test corpus are compared. For individual locations, Table 2.1 shows the mean differences of subjective attributes evaluated on 0-10 semantic differential scales by the 23 participants (n=23). Statistically significant differences in assessment distributions are computed with Wilcoxon signed rank tests [71] and shown in bold. This statistical test relies on the signs of differences between paired assessments and is non parametric. However, it ignores zero-differences in paired samples, commonly found when comparing quantities evaluated on discrete scales. Thus, Pratt's modifications of the test [72] are

further implemented to address this issue by randomly assigning a sign to zero-differences.

For all studied locations, the mean paired differences in high-level attributes is lower than 2 points. Although some statistically significant differences are found between paired assessment distributions, no consistent pattern emerges over the addressed locations. Furthermore, statistically significant differences in high-level attribute assessments can sometimes be explained by corresponding discrepancies in the perception of sound sources. For example, in the P1 location the perceived time of presence of human voices is higher for the replicated scene by 2.09 points on average. This difference is reflected in liveliness assessments for this location (-1.65 points), which is expected as both attributes are known to be correlated [1]. Similarly, the difference in perceived traffic activity in the P4 location translates to lower overall loudness and higher calmness.

In terms of perceived sound sources, high discrepancies are found that can be attributed to errors in the annotation and replication process. In the recorded scene for the P4 location, background traffic activity varies along time: it is louder in the first half of the scene than in the second half. This activity was annotated as a background source associated with a constant sound level throughout the scene. As a result, background traffic is louder for about half the duration of the replicated scene than it is in the original recording, which results in a 5.22 point increase in the perceived time of presence of traffic. A similar annotation issue for background voice activity explains the 4 points difference in the perceived time of presence of voices in the P3 location. Smaller statistically significant differences could also be attributed to the choice of isolated samples in the replication process. These samples are chosen semi-randomly according to annotations on a high-level source taxonomy (traffic, voice, and birds). In this study, for voice events no distinction is made between infant or adult speech, or between shouts and conversations. All share a common class in the isolated samples dataset, which is composed of samples from a database of recordings of read texts with generally neutral expressiveness. Similarly, traffic events are annotated regardless of the types of vehicles.

To further assess the effect of the replication process on overall perception, a principal components analysis is carried out on assessments of the five high-level attributes (resp. pleasantness, liveliness, overall loudness, interest and calmness) for the 6 recorded and 19 replicated scenes (n=25, see Section 2.3.2 for details). Individual assessments for paired recorded and replicated scenes are projected onto the resulting perceptual space. Figure 2.6 shows the distribution of projected assessments on the first two principal

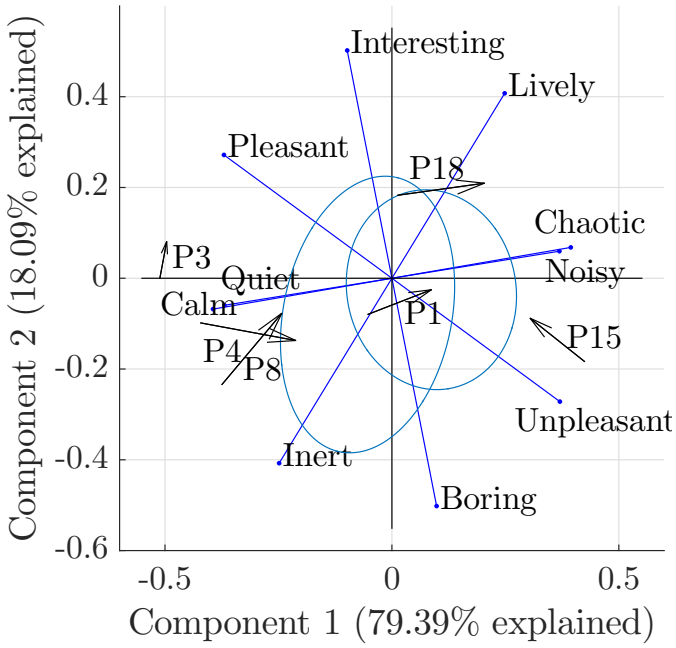


Figure 2.6: Biplot of the principal components analysis of average assessments for the 5 high-level perceptual attributes on the 6 recorded and 19 replicated scenes (n=25). Arrows indicate differences between projections of assessments for the recorded (base) and replicated (head) scenes of each location. For the P1 location ellipses show the distributions of individual assessments.

components for the P1 location as ellipses, where the center is the projected mean and axes represent standard deviations. Similar projections for the other 5 locations are available in Appendix B. For the 6 investigated locations, distributions have large intersections. This proximity confirms the perceptual likeness of replicated scenes compared to their reference recordings in diverse ambiances.

2.3.2 Scenario generation

Sound scenes simulated by *simScene*'s *generate* mode further differ from recordings as their scenarios are pseudo-randomly generated. In this case no paired comparison with recordings is possible. Assessing the effect of scenario generation on perception is thus done by studying the behaviors of perceptual attributes and potential relationships among them. To do so,

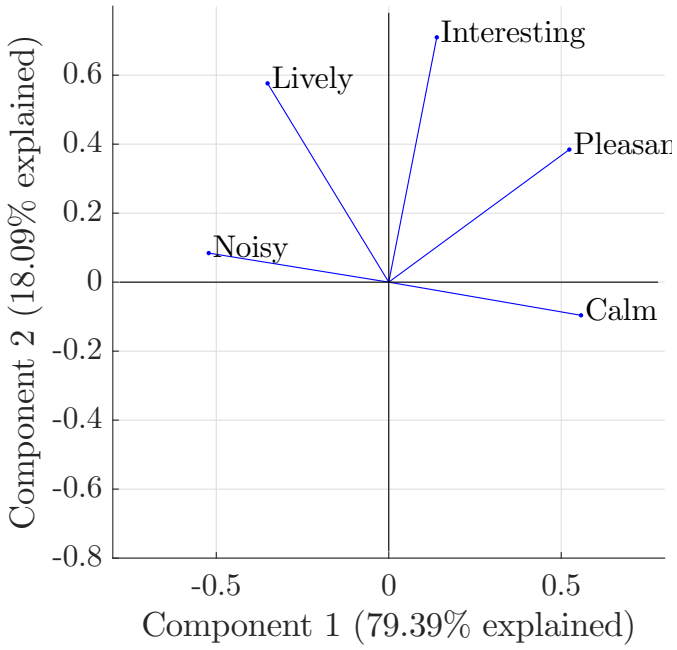


Figure 2.7: Biplot of the principal components analysis of average assessments for the 5 high-level perceptual attributes on the 6 recorded and 19 replicated scenes (n=25).

the perceptual space generated by assessments of the five high-level perceptual attributes (resp. pleasantness, liveliness, overall loudness, interest and calmness) is investigated. Following [22], the perceptual space is obtained by performing a principal components analysis (PCA) on arithmetic means of assessments over participants. Individual assessments are obtained in the range 0-10, and no standardization is applied to the data. Figure 2.7 and Figure 2.8 compare the first two components obtained for the 6 recorded and 19 replicated scenes (n=25) and the 75 simulated scenes (n=75) respectively. The perceptual space is similar for real-life and generated scenarios, although a slight rotation of the overall loudness and pleasantness axes is visible between the two spaces. The variance explained by the first two components is also similar for the two subsets, with 79.4% - 18.1% and 79.6% - 15.2% respectively. Both the distribution of attributes in the perceptual space and the variance explained by the main components are consistent with studies on perceptual dimensions in the literature [22, 23]. Assessments for individual simulated scenes are projected onto the principal components space in

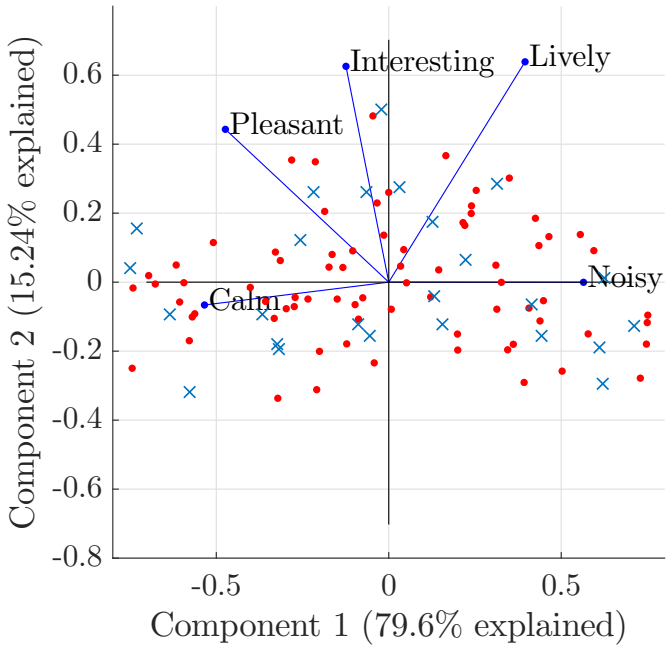


Figure 2.8: Biplot of the principal components analysis of average assessments for the 5 high-level perceptual attributes on the 75 simulated scenes (n=75). Assessments of simulated scenes (active individuals) are projected as dots, and recorded and replicated scenes (supplementary individuals) are projected as crosses.

Figure 2.8 as dots, and assessments for recorded and replicated scenes are projected as supplementary individuals and represented by crosses. The distribution of simulated scenes with generated scenarios covers that of scenes with real-life scenarios. This demonstrate the sufficient diversity of generated scenarios compared to the reference corpus of recordings.

On the whole listening test corpus, the relation between source activity descriptors and high-level attributes is further investigated. Table 2.2 shows the Pearson's correlation coefficients between perceptual attributes with assessments averaged over participants (n=100), with two-tailed significance tests. The correlations are consistent with the literature for the pleasantness, which is mainly influenced positively by birds and negatively by traffic. However, no correlation is found between pleasantness and voice activity on this corpus. This can be attributed to the nature of voice events in the isolated samples database from which the simulation process constructs acous-

Table 2.2: Pearson's correlation coefficients between perceptual attributes averaged over participants, resp. Pleasantness, Liveliness, Overall Loudness, Interest, Calmness, Time of presence of Traffic, Voices, and Birds (n=100, *: p<0.05, **: p<0.01)

	P	L	OL	I	C	$T_{T,p}$	$T_{V,p}$	$T_{B,p}$
P	1	-0.53**	-0.89**	0.66**	0.88**	-0.76**	0.05	0.57**
L		1	0.76**	0.06	-0.78**	0.17	0.60**	-0.36**
OL			1	-0.39**	-0.96**	0.59**	0.17	-0.45**
I				1	0.35**	-0.67**	0.38**	0.48**
C					1	-0.55**	-0.24*	0.48**
$T_{T,p}$						1	-0.35**	-0.42**
$T_{V,p}$							1	-0.21*
$T_{B,p}$								1

tic scenes: most extracts are recordings of read english texts (audiobooks) with overall neutral tone. The influence of voice events on pleasantness can depend on its expressiveness, with differences between shouts, laughs, conversations, adult or infant voices [33]. This distinction has less effect on the perceived liveliness of the scene, as shown by a high correlation value of 0.60 on this corpus. High correlations are also found between source activity descriptors, particularly between traffic and bird sources (-0.42). This is also found in *in situ* studies [1], and is due to the contents of specific ambiances. For example, parks or pedestrian areas contain few traffic events but high bird activity, while busy streets without vegetation may contain high traffic and low bird activity.

As discussed in Section 1.4, this study focuses on pleasantness as the main soundscape quality descriptor. In this context, the correspondence between the expression of pleasantness from sound source activity descriptors and perceptual models in the literature is verified. Multilinear regression models of pleasantness are built from the overall loudness and the time of presence of the three sources of interest. Model parameters are fitted on the arithmetic mean of assessments over participants (n=100). Each of the 15 possible combinations (with 1, 2, 3 or 4 predictors) among the 4 predictors is investigated, and for each combination a variance inflation factor (VIF) check is performed, where a combination is considered valid if all predictors verify $VIF < 5$. The criterion for selecting the best performing model is the adjusted coefficient of determination of the fitted model (R^2_{adj}). The best

model with statistically significant parameter estimates ($p < 0.05$) is:

$$\hat{P}_{1,p} = 8.99 - 0.67\ OL - 0.15\ T_{T,p} + 0.08\ T_{V,p} + 0.12\ T_{B,p} \tag{2.2}$$

where OL, $T_{T,p}$, $T_{V,p}$ and $T_{B,p}$ are the overall loudness and time of presence of traffic, voices, and birds respectively. The F-statistic of this model is $F(4, 95) = 229$ (p<0.0001) and the t-statistics (99 degrees of freedom) are 26.6, -15.5, -4.5, 2.9 and 4.8 for the intercept, OL, $T_{T,p}$, $T_{V,p}$ and $T_{B,p}$ respectively. As expected from correlation coefficients in Table 2.2, traffic and bird sources have negative and positive contributions to pleasantness respectively. However, a smaller positive contribution of voice activity is found in this model despite the absence of direct correlation to pleasantness. Overall, this expression of the pleasantness is close to perceptual models in the literature in Section 1.2, including studies relying on *in situ* questionnaires [1] and laboratory experiments with recordings of sound environments [31]. These results further confirm the adequacy of the proposed sound scene simulation method.

2.4 Indicator for the automatic annotation of simulated datasets

2.4.1 Formulation

An indicator for the automatic annotation of perceived source presence in simulated scenes is proposed within the framework of deep learning prediction discussed in Section 1.4. This indicator should perform better than other state-of-the-art acoustic indicators mentioned in Section 1.3 if it wisely considers additional information available in simulated scenes.

Source contributions for each simulated scene are available in the form of separate channels as discussed in Section 2.1. These contributions are assumed to represent the ground truth, although some extracts from the isolated samples database may also contain other sources or background noise as they are recorded in real-life environments. Finely annotated corpora of simulated scenes were previously introduced in environmental sound event detection challenges [73]. The corresponding deep learning task consists in predicting event onsets and offsets with precision in tens of milliseconds. Physical source presence is trivially annotated in this case: the source is considered present at time t if the energy around time t is greater than zero in the corresponding channel.

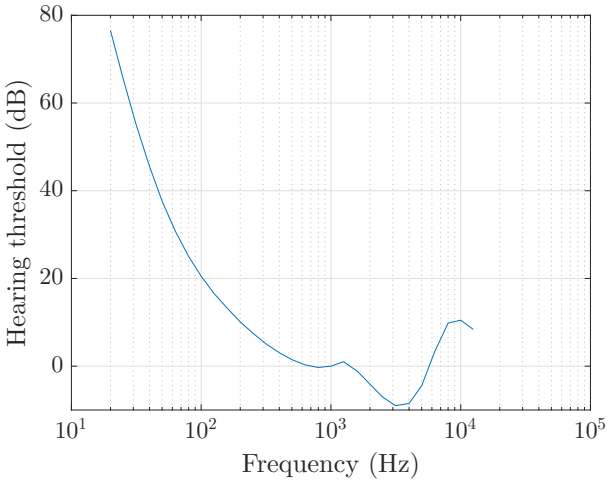


Figure 2.9: Lowest equal-loudness contour (dB SPL) in the ISO 226:2003 norm, taken as an absolute threshold of hearing curve.

However, the task investigated in the present study differs from this paradigm and addresses perceptual presence instead. As discussed in Section 1.4, this notion only makes sense at longer time scales of about 1 s. In addition, perceptual presence implies that the source is heard within the mixture. First, this may not be the case if its sound level is too low. The lowest equal-loudness contour defined in the ISO 226:2003 norm [74] and shown in Figure 2.9 is thus regarded as a hearing threshold and applied to each source contribution independently, with the equivalent sound level associated to the scene as reference. This hearing threshold curve is not fully accurate as it assumes harmonic signals, but it provides a lower bound in case of sound objects with full-band energy. Second, in polyphonic environments a source s may be masked by other events simultaneously occurring. Accounting for interactions between individual sound sources in the mix is thus necessary.

This problem is related to the characterization of mechanisms of auditory attention and the salience of sound events within an environment, which have been the subject of extensive research [75, 76]. Typical approaches view a sound scene as an ensemble of auditory streams with defining acoustical properties (e.g. different sources). The attention switches between streams over time through bottom-up and top-down mechanisms, respectively based on the saliency of auditory streams and voluntary attention. The authors of [77] propose a method for computing saliency maps of sounds from time-

frequency amplitudes and variations (contrast). In [78], time-dependant saliency scores are obtained by applying binary time-frequency masks corresponding to auditory streams to this saliency map of the mixed sound scene. The saliency scores are input to a model simulating auditory attention switching. In [79] a more complex saliency map is proposed that specifically exploits the temporal properties of sounds, with features such as the envelope, pitch and bandwidth. The authors of [80] apply a similar approach to characterize attention for two sound scenes heard simultaneously.

Although attention mechanisms naturally occur in listening experiments, the concept of perceptual source presence investigated here is not explicitly constrained to a single active sound object at a given time. Relevant time-dependant saliency scores could still be derived from separate streams of sound source contributions. However, for the purpose of simplicity and control, this study instead addresses a simplified context of active listening by the passer-by of the sound environment. A new indicator is thus developed that solely addresses the audibility of components in polyphonic scenes. Specifically, the indicator relies on the local emergence of sound sources to roughly approximate auditory masking effects.

Consider a polyphonic sound object composed of an harmonic at 125 Hz and a stationary noise component. Figure 2.10 illustrates the spectral content of both components for white and pink noise and for signal to noise ratios (SNR), *i.e.* emergence of the harmonic signal, of -15 dB and 15 dB. In cases (b) and (d) where the SNR is 15 dB, both components can be heard in the mix. However, when the SNR is -15 dB the harmonic is heard with white noise (a) but not with pink noise (c). This points to limitations in the capacity of the full band emergence to describe the perceived presence of a source, which compares sound levels aggregated over the range of audible frequencies.

A real-life example of this phenomenon is the simultaneous activity of traffic and bird sources. If the traffic source has low energy in high frequencies, the bird source will likely be heard despite low full-band emergence. Conversely, the same traffic source with an added high-frequency component, for instance braking sounds, can mask the bird source. A source masking model should thus be defined from the emergence in individual frequency bands as opposed to the full-band emergence. However, using emergences on high-resolution fine-band spectra is not straightforward, as it requires accounting for critical bands where a locally harmonic signal masks components in surrounding frequencies [81]. Here, an indicator based on the third-octave emergence is instead proposed. The emergence of source s is

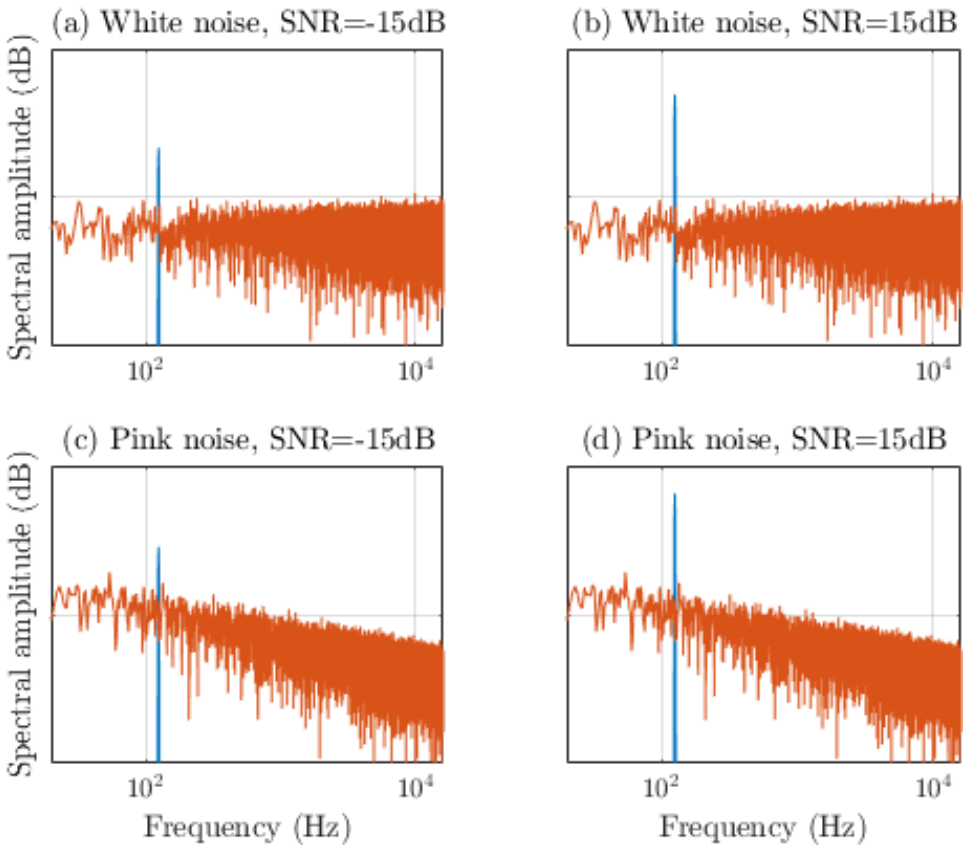


Figure 2.10: Spectra of harmonic and noise components in polyphonic signals for different signal to noise ratios and noise colors. The harmonic component is perceptually masked only in (c).

defined as:

$$\Delta L_s\left(t,f\right) = L_s\left(t,f\right) - L_{\bar{s}}\left(t,f\right) \tag{2.3}$$

where L_s is the sound level of source s, $L_{\bar{s}}$ is the sound level of all other sources combined by adding the corresponding channels in the time domain, and t and f refer to 1 s frames and third octave spaced frequency bands respectively. First, third octave bands where the source of interest s is emergent are selected by applying a threshold α on $\Delta L_s\left(t,f\right)$, under which the source is considered as completely masked:

$$\mathbb{1}_{\Delta L_s(t,f)>\alpha} \tag{2.4}$$

where $\mathbb{1}$ denotes the indicator function, *i.e.* $\mathbb{1}_{\Delta L_s(t,f)>\alpha}$ returns 1 if $\Delta L_s(t,f) >$

α and 0 otherwise. Then, on a given 1 s time frame the source is considered present if the average emergence of third octave bands selected by eq.2.4 is greater than a threshold value β:

$$\mathbb{1}\left[\frac{\sum_{f=1}^{N_f}\Delta L_s(t,f)\mathbb{1}_{\Delta L_s(t,f)>\alpha}}{\sum_{f=1}^{N_f}\mathbb{1}_{\Delta L_s(t,f)>\alpha}}>\beta\right] \tag{2.5}$$

where N_f is the number of third octave bands. It is expected that β should be greater than α, thus more restrictive, as otherwise eq. 2.5 returns 1 if eq. 2.4 returns 1 for at least one third octave band. Note that the average is taken over logarithmic sound levels, thus it cannot be associated to a full band sound level.

This binary source presence indicator annotates 1 s frames in simulated scenes to train a deep learning model in Chapter 3. However, an estimation of the time of presence is necessary as part of models of high-level perceptual attributes. Here, the estimated time of presence for source s, designated $\hat{T}_s(\alpha,\beta)$, is obtained by averaging presence labels given by eq.2.5 over time:

$$\hat{T}_s(\alpha,\beta)=\frac{1}{N_t}\sum_{t=1}^{N_t}\mathbb{1}\left[\frac{\sum_{f=1}^{N_f}\Delta L_s(t,f)\mathbb{1}_{\Delta L_s(t,f)>\alpha}}{\sum_{f=1}^{N_f}\mathbb{1}_{\Delta L_s(t,f)>\alpha}}>\beta\right] \tag{2.6}$$

where N_t is the number of 1 s frames in the sound scene.

The α and β parameters of the model are optimized so that the correlation between $\hat{T}_s(\alpha,\beta)$ and the perceived time of presence $T_{s,p}$ is maximized. To do so, the optimization criterion is the average Pearson correlation coefficient r between both variables averaged over all $N_s=3$ sources:

$$\alpha_{opt},\beta_{opt}=\underset{\alpha,\beta}{\arg\max}\frac{1}{N_s}\sum_{s=1}^{N_s}r\left(T_{s,p},\hat{T}_s(\alpha,\beta)\right) \tag{2.7}$$

Optimal values of α and β are found via grid search with $\alpha,\beta\in[-20dB,10dB]$ by steps of 1 dB. On the 19 replicated scenes and 75 simulated scenes of the listening test corpus annotated in Section 2.2.3, optimal values are found as $\alpha=-14dB$ and $\beta=-7dB$. As expected, β is greater than α, which justifies the interest two separate thresholds compared to only α. The same optimisation process was previously carried out on a similar dataset in [61], with different optimal values of $\alpha=-6dB$ and $\beta=-5dB$. However, the optimisation criterion is found to be very stable to changes in both parameters. This is illustrated by Figure 2.11 on the current listening test corpus.

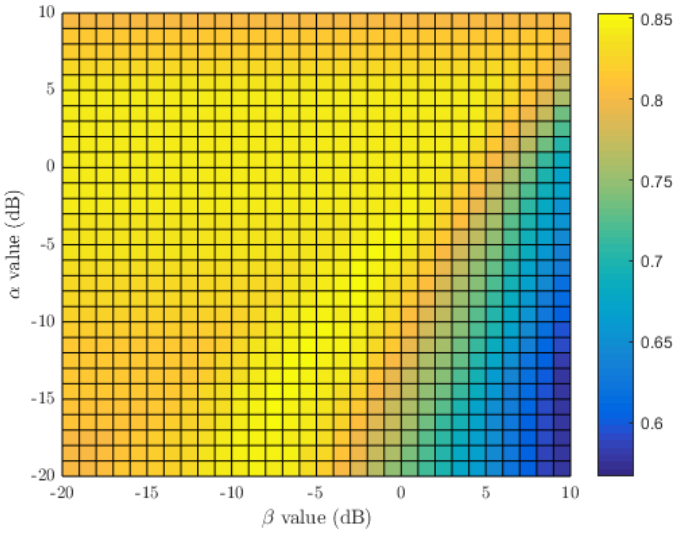


Figure 2.11: Average Pearson correlation coefficient between the proposed time of presence estimation $\hat{T}_s(\alpha, \beta)$ and subjective annotations as a function of α and β values.

The proposed source presence indicator suffers from limitations due to its simplicity. First, the α emergence is applied independently on each time frame and frequency band. Thus, spectral and temporal masking effects where a component with locally high energy can mask other components in surrounding frequencies and time frames respectively, are not explicitly taken into account. In the case of temporal masking, this issue is attenuated on third octave representations: temporal masking typically occurs on scales shorter than 100 ms, whereas the proposed indicator is defined on frames of 1 s of audio. Similarly, third octave bands attenuate the influence of spectral masking by averaging information over large bandwidths, particularly in high frequencies. Also, the proposed approach only models energetic masking as it is based on sound level differences, whereas informational masking, *i.e.* the masking of information within a source by a simultaneously active source with similar components, is not represented.

Second, time of presence estimations $\hat{T}_s(\alpha, \beta)$ are obtained by averaging presence labels over time. This assumes that the impact of a sound event on the perceived time of presence does not depend on its time of occurrence, and that no interaction exists between subsequent events from different sources. A more complete model would estimate the time of presence as a weighted average of presence labels with varying weights along time. Alternatively, a

recurrent process could better account for the impact of the time of occurence of events, both independently and in the context of other source activity within the scene. Here, the two parameters of the proposed indicator are optimized on a small corpus of 94 sound scenes. The number of parameters is thus deliberately limited to avoid overfitting on the available subjective annotations. However, because of these simplifications, the validity of the proposed indicator with found optimal parameter values is not ensured for different scene duration. The scope of validity of this study is therefore limited to 45 s sound scenes.

2.4.2 Evaluation

The performance of the proposed time of presence indicator $\hat{T}_s(\alpha, \beta)$ is first evaluated by its relation to the perceived time of presence for traffic, voice and bird sources, as well as high-level perceptual attributes annotated in Section 2.2.3. To do so, it is compared to state-of-the-art acoustic indicators presented in Section 1.3. The following acoustic indicators recurring in monitoring applications or models of soundscape quality are computed on from the mixed sound scene using the Matlab ITA-toolbox [82]:

- 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, in dB. These indicators are associated to the sound level of emergent events, the overall sound level, and the sound level of background sources respectively.
- LA_{50}: 50th percentile of the A-weighted sound level in dBA, as an alternative to the L_{50}.
- $L_{50,1kHz}$: 50th percentile of the Z-weighted sound level for the 1 kHz frequency band in dB, also associated to the overall sound level of the sound scene.
- $LA_{10} - LA_{90}$: Emergence indicator in dBA, included in the pleasantness prediction model presented in [31].
- $TFSD_{4kHz(1/8s)}$ and $TFSD_{500Hz,1s}$: Time and Frequency Second Derivative proposed in [1] as descriptors of bird and voice activity respectively.

In the case of simulated scenes, ground truth source contributions are outputs of the generation process. Additional source-specific indicators are derived from this information:

- $L_{eq,s}$: Equivalent sound level for source s in dB.
- ΔL_s: Full-band emergence of source s in dB, taken as the difference between the sound level of source s and that of all other sources.

The s subscript in these indicators, as well as in the proposed $\hat{T}_s(\alpha, \beta)$ time of presence estimation is replaced with T, V or B for traffic, voice and bird sources respectively.

Table 2.3: Pearson's correlation coefficients between physical and perceptual (resp. Pleasantness, Liveliness, Overall Loudness, Interest, Calmness, Time of presence of Traffic, Voices, and Birds) indicators (n = 92).

	P	L	OL	I	C	$T_{T,p}$	$T_{V,p}$	$T_{B,p}$
LA_{eq}	-0.86**	0.68**	0.92**	-0.37**	-0.88**	0.66**	0.07	-0.41**
LA_{50}	-0.84**	0.67**	0.91**	-0.33**	-0.87**	0.63**	0.06	-0.35**
L_{eq}	-0.88**	0.67**	0.91**	-0.44**	-0.88**	0.71**	0.06	-0.46**
L_{10}	-0.87**	0.65**	0.90**	-0.44**	-0.86**	0.71**	0.06	-0.47**
L_{50}	-0.89**	0.65**	0.92**	-0.43**	-0.89**	0.71**	0.03	-0.44**
L_{90}	-0.86**	0.68**	0.92**	-0.39**	-0.89**	0.67**	0.07	-0.40**
$L_{50,1kHz}$	-0.88**	0.69**	0.92**	-0.42**	-0.89**	0.73**	0.08	-0.50**
$L_{10} - L_{90}$	0.13	-0.18	-0.24*	-0.06	-0.22*	-0.01	-0.01	-0.09
$TFSD_{500Hz,1s}$	0.07	0.41**	0.11	0.28**	-0.15	-0.39**	0.74**	-0.17
$TFSD_{4kHz,1/8s}$	0.52**	-0.43**	-0.49**	0.41**	0.52**	-0.54**	-0.18	0.63**
$L_{eq,T}$	-0.58**	0.20	0.46**	-0.46**	-0.42**	0.71**	-0.16	-0.36**
$L_{eq,V}$	-0.17	0.50**	0.31**	0.08	-0.37**	-0.04	0.71**	-0.40**
$L_{eq,B}$	0.27*	-0.04	-0.11	0.35**	0.18	-0.24*	-0.04	0.71**
ΔL_T	-0.45**	-0.11	0.26*	-0.59**	-0.22*	0.66**	-0.51**	-0.26*
ΔL_V	0.04	0.50**	0.17	0.35**	-0.20	-0.38**	0.59**	-0.01
ΔL_B	0.21*	-0.25*	-0.26*	0.08	0.25*	-0.25*	-0.10	-0.03
$\hat{T}_T(\alpha_{opt}, \beta_{opt})$	-0.53**	-0.05	0.35**	-0.57**	-0.29**	0.81**	-0.39**	-0.37**
$\hat{T}_V(\alpha_{opt}, \beta_{opt})$	0.12	0.44**	0.05	0.35**	-0.11	-0.39**	0.81**	-0.16
$\hat{T}_B(\alpha_{opt}, \beta_{opt})$	0.56**	-0.30**	-0.46**	0.55**	0.51**	-0.57**	-0.08	0.91**

Table 2.3 shows the Pearson correlation coefficients between acoustic indicators and the arithmetic mean of perceptual assessments on the listening test corpus, with two-tailed significance tests (*:p<0.05, **:p<0.01). Ground truth source contributions are unknown for the 6 recorded scenes, thus the proposed indicator $\hat{T}_s(\alpha, \beta)$, the source sound level $L_{eq,s}$ and emergence ΔL_s cannot be computed and these scenes are removed from the study for fair comparison. Additionally, two simulated scenes where only one of the three sources of interest is active yield infinite emergence values. These scenes are removed from the study to ensure numerical stability of emergence indicators. As a result, this analysis is conducted on $n = 92$ sound scenes among the 100 in the listening test corpus.

Even if it considers each source independently, the source-specific sound level $L_{eq,s}$ correlates consistently well with the perceived time of presence for the three sources of interest (r=0.71). The full band emergence ΔL_s however fails to describe the perceived bird activity (r=-0.03). This is consistent with observations in Section 2.4.1 where masking depends on the spectral distribution of energy. The proposed indicator $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ is correlated to corresponding subjective assessments for the three sources, with r=0.81, r=0.81 and r=0.91 for traffic, voice and bird time of presence respectively. It also performs better than the state-of-the-art descriptors of voice and bird activity, $TFSD_{500Hz,1s}$ (r=0.74) and $TFSD_{4kHz,1/8s}$ (r=0.63) respectively. The $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ indicator correlates poorly with the perceived time of presence of other sources, indicating good discriminative properties similarly to the TFSD variants. High negative correlation between the subjective activity of traffic and the predicted activity of bird sources (r=-0.57) is attributed to the contents of sound scenes from different ambiances as discussed in Section 2.3.2. Correlations between $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ and high-level attributes are also expected, for example high negative (r=-0.53) and positive (r=0.56) contributions of traffic and bird sources to pleasantness are found, and liveliness is mainly correlated to human voice activity (r=0.44).

In the framework described in Section 1.4, the deep learning model is trained to predict source presence resulting in the estimation of the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ indicator. These estimations are then input to a linear model in order to obtain the pleasantness of the sound scene. This model can first be found by direct optimisation over the assessments on the listening test corpus. To do so, a multilinear regression model of pleasantness is constructed. Existing perceptual models in Section 1.2 include the perceived overall loudness in addition to the time of presence of traffic, voice and bird sources. In Table 2.3, all overall sound level indicators achieve similar performance (r>0.9) in describing the perceived overall loudness of the sound scene. The L_{50} is the best performing indicator, and is taken as a predictor. The three other predictors are composed of $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ for the three sources of interest. The variance inflation factor (VIF) is computed for all combinations of predictors and only those verifying $VIF < 5$ are included in the same model to avoid collinearities. On n=92 sound scenes, the best model obtained in terms of R^2_{adj}, *i.e.* adjusted coefficient of determination, is:

$$\hat{P}_{1,\varphi} = 16.74 - 0.18\ L_{50} + 1.01\ \hat{T}_B(\alpha_{opt}, \beta_{opt}) \tag{2.8}$$

where φ indicates a model from acoustic indicators. The F-statistic of the $\hat{P}_{1,\varphi}$ model is $F(2, 89) = 210$ (p<0.0001) and the t-statistics (91 degrees of

Table 2.4: Performance of baseline models for pleasantness prediction.

	$RMSE$	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**

**: p<0.01

freedom) are 21.6, -15.8, and 3.8 for the intercept, L_{50} and $\hat{T}_B(\alpha_{opt}, \beta_{opt})$ coefficient estimates respectively. Contrary to perceptual models in Section 2.3.2 and the literature, traffic and voice contributions as described by $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ are not found to contribute significantly to the prediction of pleasantness for this corpus. The absence of a traffic descriptor is attributed to high correlations between the L_{50} and the perceived time of presence of traffic $T_{T,p}$ in Table 2.3 (r=0.71). A possible explanation is the construction of simulated scenes, which only contain the three sources of interest. Traffic activity thus contributes more to the overall sound level than in real life situations: quiet ambiances such as parks and pedestrian streets are less likely to have continuous traffic, while high sound levels in busy streets are always due to traffic. Thus, the absence of other typically loud sources to the taxonomy, such as construction work and other forms of transportation, increases the correlation between sound level and traffic activity. Voice activity is not significantly correlated to pleasantness in Table 2.3 (r=0.12), resulting in its absence in the $\hat{P}_{1,\varphi}$ model. This can also be explained by the limited diversity of voice event sources in the simulated scene corpus: voice events are primarily composed of recordings of read English with few different speakers and mostly neutral expressiveness.

The $\hat{P}_{1,\varphi}$ model is compared to baselines proposed in [31] and in [1], referred to as $\hat{P}_{2,\varphi}$ and $\hat{P}_{3,\varphi}$ respectively. The $\hat{P}_{2,\varphi}$ model does not explicitly involve specific sound sources and instead considers the $L_{10} - L_{90}$ to describe overall event emergence. The $TFSD_{500Hz,1s}$ and $TFSD_{4kHz,1/8s}$ appear in the $\hat{P}_{3,\varphi}$ as descriptors of voice and bird sources. Coefficients for both models are re-optimized on the studied data for a fair comparison:

$$\hat{P}_{2,\varphi} = 18.67 - 0.20\ L_{50} - 0.02\ (L_{10} - L_{90}) \tag{2.9}$$

$$\hat{P}_{3,\varphi} = 30.18 - 0.16\ L_{50,1kHz} + 8.92\ TFSD_{500Hz,1s} + 2.99\ TFSD_{4kHz,1/8s} \tag{2.10}$$

On the listening test corpus the $L_{10} - L_{90}$ emergence in $\hat{P}_{2,\varphi}$ does not contribute significantly to the model ($p>0.05$), and the same overall performance metrics are obtained with only the L_{50}. Table 2.4 summarizes the performance of compared models from acoustic indicators, compared to the perceptual model found in Section 2.3.2. As expected, the perceptual model $\hat{P}_{1,p}$ yields the best prediction performance with a root mean squared error of 0.61. However, the error of other models remains 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 $\hat{P}_{3,\varphi}$, although its expression is likely a result of the specific corpus of this study. Its relevance for pleasantness prediction in real-life environments should thus be validated on a corpus of more diverse sound scenes, in terms of ambiances and source taxonomy.

Alternatively, proposed indicators can be substituted in a state-of-the-art perceptual model of pleasantness. In this case, estimations of the perceived time of presence are obtained by applying an affine transformation to the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ indicator independently for the three sources. Optimizing these transformations on the listening test corpus yields:

$$\hat{T}_{T,p} = 6.65\ \hat{T}_T(\alpha_{opt}, \beta_{opt}) \tag{2.11}$$

$$\hat{T}_{V,p} = 6.46\ \hat{T}_V(\alpha_{opt}, \beta_{opt}) + 1.64 \tag{2.12}$$

$$\hat{T}_{B,p} = 7.15\ \hat{T}_B(\alpha_{opt}, \beta_{opt}) + 2.58 \tag{2.13}$$

$$\hat{OL} = 0.18\ L_{50} - 3.52 \tag{2.14}$$

$$\hat{P}_{4,\varphi} = 7.11 - 0.38\ \hat{OL} - 0.14\ \hat{T}_{T,p} + 0.20\ \hat{T}_{V,p} + 0.15\ \hat{T}_{B,p} \tag{2.15}$$

where coefficients in eq. 2.15 are not optimized, but taken from the perceptual pleasantness model proposed in [31] instead. However, without re-optimization of the perceptual model's coefficients the performance of this model is poor on the listening test corpus, on all evaluated metrics ($RMSE = 1.22$, $R^2_{adj} = 0.62$, $r = 0.80$) compared to all models with optimized coefficients in Table 2.4. Replacing coefficients in $\hat{P}_{4,\varphi}$ by those of the perceptual model optimized on the listening test corpus in eq. 2.2 with the same predictors yields $RMSE = 0.96$, $R^2_{adj} = 0.75$ and $r = 0.87$. In terms of pleasantness prediction performance this model is closer to baseline models from acoustic indicators, but still yields slightly higher prediction errors.

Chapter conclusion

Constructing a database of isolated samples of sources of interest as well as distributions of source activity extracted from annotated recordings allows the generation of large datasets with diverse scenarios. The perceptual

properties of simulated corpora, including the space induced by high-level perceptual attributes and their relation to scene content, match those of recordings. Simulated scenes can be annotated with labels of source presence that correlate well with subjective assessments using the proposed indicator. This allows the creation of large simulated corpora labeled for the perceived source presence prediction task, with sufficient total duration and diversity of scenarios and polyphonies to train predictive deep learning architectures. In Chapter 3, these results lead to the design and training of deep learning architectures on the target task.

Chapter 3

Prediction of the perceived time of presence from sensor measurements using deep learning

Two deep learning architectures are proposed to predict labels of perceived source presence from spectral representations measured by acoustic sensors. These architectures rely on a frame-independent or recurrent decision processes respectively to produce predictions. Both models operate at faster than real-time speeds, and can thus be applied to sensor networks in continuous monitoring applications.
Evaluation is conducted on a large dataset of simulated scenes automatically annotated in terms of perceived source presence. In this setting, experiments show that the proposed models predict presence labels with satisfying accuracy.

3.1 Introduction

The relevance of deep learning models is investigated for the task of predicting the perceived presence of sources in the applicative context of continuous

monitoring with sensor network measurements, and specifically in the Cense project. Data from the Cense sensor network implemented in Lorient is composed of third-octave spectra with 29 frequency bands in the $[20Hz, 20kHz]$ range and information aggregated on 125 ms (fast) temporal frames. On the other hand, the indicator proposed in Section 2.4.1 can automatically annotate audio frames of 1 s in terms of binary (absent-present) traffic, voices, and birds sources presence. The inputs and outputs of the task are thus constrained to textures of sensor measurements over 1 s, *i.e.* 8 frames of consecutive fast third-octave measurements, associated with presence labels. The presence prediction task is formulated as a multilabel classification problem where any number of sources within the taxonomy can be active simultaneously.

To the best of our knowledge, this task has not been specifically investigated in the literature. However, it is related to several tasks actively studied by the detection and classification of acoustic scenes and events (DCASE) community[1]:

- Environmental sound classification (ESC), which consists in categorizing short monophonic excerpts of sound events within a taxonomy.
- Acoustic scene classification (ASC), which consists in classifying longer extracts as ambiances or environments.
- Sound event detection (SED), which consists in predicting the onset-offsets of specific sound events within monophonic or polyphonic sound scenes.

Although the scopes of these tasks are different, they all require extracting information about the contents of the audio signal in terms of active sound sources and can thus be formulated similarly. Early approaches rely on traditional machine learning tools such as support vector machines [83] or Gaussian mixture models [84] together with carefully selected representations of the audio signal. These methods are outperformed by recent deep learning approaches, with convolutional and recurrent networks as the most popular model architectures [85]. Convolutional neural networks were first proposed in the image processing community [86] to extract visual patterns from images using groups of translation-invariant filters, in order to perform object recognition. In environmental sounds, sources are often characterized by time-frequency patterns in spectral representations of the signal. These patterns can thus be efficiently identified by convolutional networks, leading

[1] https://dcase.community/

to state-of-the-art performances in classification tasks [87, 15, 88]. Recurrent neural networks regard the signal as a sequence of short frames, and propagate information along time by updating a recurrent hidden state to draw local predictions based on short-term and long-term context. Recurrent architectures perform particularly well in event detection tasks [89, 90].

In terms of input audio signal representations, most approaches process spectral transforms such as linear magnitude spectrograms, Mel spectrograms or Mel frequency cepstral coefficients (MFCC). These representations are computed on very short time scales, typically tens of milliseconds with overlapping frames, and with sufficient frequency resolution to capture modulations characterizing all sources of interest. The multilabel classification task investigated here differs from the literature in this regard, as the input representation is constrained to that measured by sensors of the Cense network. These measurements are third-octave energies aggregated on 125 ms non-overlapping frames. Sources characterized by modulations on scales shorter than 125 ms or narrow high-frequency bands, such as birdsong, may thus be difficult to identify. However, the current task does not introduce constraints on the architecture of the deep learning model, which can be inspired from existing approaches for classification or event detection.

3.2 Controlled dataset

A dataset is constructed for the purpose of training and evaluating deep learning models on the prediction of the perceived presence of traffic, voices, and birds sources. To allow automatic annotation with the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ indicator proposed in Section 2.4.1, this dataset is composed of sound scenes simulated by *simScene*'s *generate* mode. Typical datasets for related tasks range from a few hours to tens of hours in duration depending on the task [91, 92, 93, 56].

The deep learning dataset is composed of a development subset and an evaluation subset. The development set is generally split into a training and a validation subsets, respectively to optimize model parameters and to monitor convergence and overfitting behaviors. The development set contains 400 sound scenes of 45 s each for a total duration of 5 h. The scenes are equally distributed over the five ambiances (quiet street, noisy street, very noisy street, park, square) in the taxonomy. Model performance is assessed after training on the independent evaluation set, which contains 200 sound scenes of 45 s each (total duration of 2.5 h).

The scene simulation process for this dataset retains the taxonomy of

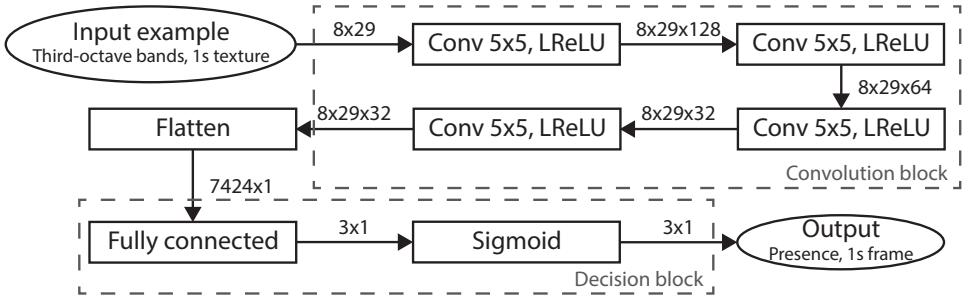


Figure 3.1: Architecture of the proposed convolutional neural network for source presence prediction in 1 s textures of third-octave fast measurements.

ambiances and sound sources described in Section 2.2.2, as well as source activity distributions extracted from the corpus of recordings in the GRAFIC project from which scenarios are generated. The isolated samples database from which scenes are constructed is different from that of the listening test corpus, and contains 12 min 12mn, 49 min, and 73 min of background traffic, voice and bird extracts respectively, and 32 min, 12 min, and 5 min of event traffic, voice and bird extracts respectively. Yet, the origin of samples is similar: all voice events are extracts of read English texts from the Librispeech dataset, and other sound sources are represented by recordings selected in the Freesound repository. Furthermore, the development and evaluation subsets must be independent to ensure that metrics computed during evaluation reflect the generalisation capabilities of the model. Thus, the isolated samples database is split in two parts. Two-thirds of available extracts for each source are dedicated to simulating scenes in the development set only, and evaluation scenes contain samples from the remaining one-third of available extracts.

3.3 Architectures

3.3.1 Convolutional neural network

In the applicative context of continuous source presence prediction from acoustic sensor network data, the proposed model should be limited in terms of computational and memory complexity. In addition to their capacity at extracting time-frequency patterns from audio spectra, convolutional neural networks are uniquely suited to this purpose as filters in convolution layers can be applied in parallel.

The architecture of the proposed model is shown in Figure 3.1. It is composed of four convolution layers with common kernel size 5x5 and 128, 64, 32 and 32 output channels respectively, and a fully connected layer with output size 3. Each example processed by the model is an 8x29 input matrix of third-octave measurements, associated to three ground truth output values corresponding to the absence (0) or presence (1) of each source of interest in the input spectrum. The inputs are not normalized so that the model may include sound level information in the inference process.

The role of convolution layers is to extract time-frequency patterns from the input representation that are relevant to solving the presence prediction task, *i.e.* that characterize specific sound sources. The first convolution layer applies a filterbank of 128 filters to the input third-octave spectrum. The receptive field of filters is given by their kernel size of 5x5, thus filters in the first layer have a receptive field of 5 temporal frames (625 ms) and 5 third-octave bands. Here, no stride is applied during convolutions and the input is zero-padded so that the layer's output is of the same dimension (8x29), *i.e.* convolutions do not change the sampling rate of the input representation. The output of the first convolution layer is thus composed of 128 filtered spectra referred to as channels. In following layers, for an input representation with M channels and an output representation with N channels, the output is obtained by applying N independent groups of filters to the input. Each group contains M filters each applied to a different input channel, and a single output channel is obtained by summing the results. Formally, for input channels h_m^{i-1} and filters k_{mn}^i, output channels h_n^i are computed as:

$$h_n^i = \sum_{m=1}^{M} k_{mn}^i * h_m^{i-1}, n = 1, .., N \tag{3.1}$$

Each subsequent convolution layer with kernel size K increases the receptive field of the model by $K - 1$ in time and frequency. With 5 layers the total receptive field of the convolution block is thus 17x17, although it is limited in practice by the input size to 8x17, as examples only span 8 frames (1 s) of signal. Each of the four convolution layers is followed by a leaky rectified linear unit (LeakyReLU) activation of expression $max(\alpha x, x)$ applied element-wise to the hidden states [94]. The α parameter is constant and equals 0.01 in all experiments.

The output of the last convolution layer is flattened into a vector of dimension 7424 and input to a fully connected layer. This final layer is a linear projection from $\mathbb{R}^{7424}$ to $\mathbb{R}^3$ parametrized by a matrix W of dimension 3x7424 and a bias term of dimension 3. The role of this projection is to

summarize information in the time, frequency and channel dimensions of the convolution block's output to predict three scalar values, each corresponding to a source of interest (resp. traffic, voices, and birds). A sigmoid activation is further applied to the output values to obtain predictions in the $]0, 1[$ range, which can be interpreted as probabilities for each source to be present in the input representation. Powerful loss functions based on the negative log-likelihood then guide the training process of the model (See Section 3.3.3 for details). During evaluation, binary decisions on the presence of sources are obtained by applying a threshold of 0.5 to each output value independently. The corresponding source is considered present (1) on the 1 s time frame of the input if the model outputs a value greater than 0.5, and absent (0) if the output is less than 0.5. The model is fully deterministic and contains a total of about 300 000 parameters.

3.3.2 Recurrent decision process

The convolutional neural network presented in Section 3.3.1 predicts the presence of sources of interest on independent 1 s textures of acoustic data. However, some event sound sources are characterized by activity on longer time scales, particularly traffic-related events. Accounting for information of past context in the decision process may thus help identify these events. Recurrent neural networks are specifically designed to model this type of sequential relationships. In a simple recurrent layer, information is propagated along time by adding a parametric recurrent connection to a fully connected layer:

$$h_t^{(i)} = f(\mathbf{W_r^{(i)}} \mathbf{h_{t-1}^{(i)}} + W^{(i)} h_t^{(i-1)} + b^{(i)}) \tag{3.2}$$

where $h_t^{(i)}$ is the hidden state of layer i at timestep t, W_r^i is the projection matrix of the recurrent connection, W^i is the projection matrix of the input connection, f is a nonlinear activation function and b^i is a learned bias term. At each timestep, a new element of the input sequence is introduced to the recurrent layer in order to update the hidden state, which is then processed to produce an output for that timestep. Backpropagation through time of the gradient during training is achieved by unrolling the network, that is computing gradients of the loss function with respect to model parameters at each timestep independently, then summarizing contributions along time to obtain a single update per parameter.

Given that the transformation from $t-1$ to t in the recurrent layer is a linear projection defined by a matrix W_r, backpropagating one timestep multiplies the gradient at time t by W_r to obtain the gradient at time $t-1$.

Backpropagating over a sequence of T timesteps thus implies that the gradient is multiplied by W_r T times, allowing recurrent cells to capture dependencies in long sequences is limited [95]. If the recurrent projection matrix W_r contains small weights, the gradient in early timesteps becomes too small compared to that of last timesteps to have a significant impact on parameter updates. This behavior is known as vanishing gradient and causes difficulties in learning long-term dependencies in data sequences. Conversely, if W_r contains large weights, the gradient in early timesteps explodes in value. As a result the learning process may become unstable and model parameters may fail to converge. The potential for vanishing or exploding gradients within a recurrent model is formally measured by the spectral radius of the recurrent projection matrix W_r, defined as the largest of its eigenvalues λ. If $max_i|\lambda_i| < 1$ the model is prone to vanishing gradients, and $max_i|\lambda_i| > 1$ indicates a risk of exploding gradients, with both behaviors occuring exponentially in $O(n^T)$ where T is the total duration of the sequence.

The Long Short-Term Memory (LSTM) [96] is an alternative to the recurrent cell that solves vanishing gradients. To do so, it introduces a second hidden state that is never directly transformed by a projection. Instead, the output hidden states of the LSTM cell at time t are an additive combination of the hidden states at time $t-1$ and the current input. Information is filtered by three parametric gates (resp. input, forget, output). However, because of the three gates the LSTM cell contains four times as many parameters as a normal recurrent cell with the same hidden state dimension, and memory usage is further increased by the additional hidden state. These factors result in a significantly higher computational complexity, which makes the LSTM unsuitable in the current application.

The gated recurrent unit (GRU) [97] is introduced as a simplification of the LSTM cell that conserves its gradient backpropagation stability. As shown in Figure 3.2, the GRU cell only contains one hidden state and two parametric gates controlling the propagation of information from the current input (reset gate r_t) and the combination of the current input with past information (update gate u_t):

$$u_t = \sigma\left(W_u \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_u\right) \tag{3.3}$$

$$r_t = \sigma\left(W_r \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_r\right) \tag{3.4}$$

$$h_t = u_t \odot h_{t-1} + (1 - u_t) \odot f_W\left(h_{t-1}, r_t \odot x_t\right) \tag{3.5}$$

where σ is the elementwise sigmoid function, h_t and x_t are the hidden state

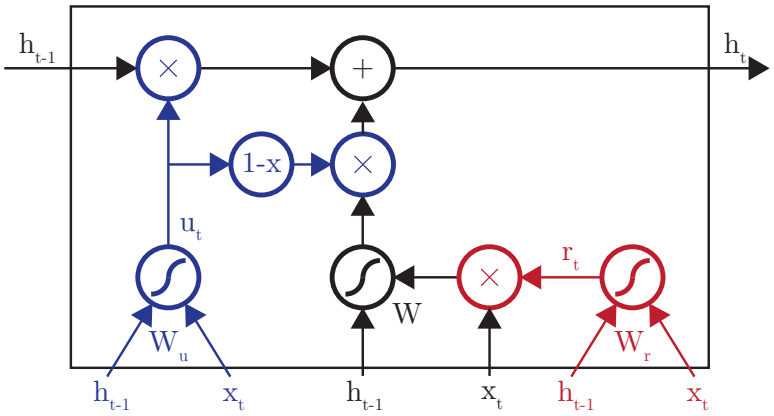


Figure 3.2: Information flow in a gated recurrent unit cell. The update (blue) and reset (red) gates filter information from the current input x_t and recurrent state h_{t-1} to compute the hidden state h_t.

and the input of the layer at time t respectively, and $\odot$ denotes an element-wise multiplication.

The architecture of the proposed convolutional neural network with recurrent decision using a gated recurrent unit is shown in Figure 3.3. The network processes a sequence of T textures of dimension 8x29 corresponding to 1 s of third-octave measurements, and outputs T predictions of source presence, *i.e.* one for each timestep. Two consecutive textures in the input sequence contain overlapping information for 875 ms as illustrated in Figure 3.4. At each timestep of the sequence, the model extracts information from the input 1 s texture using the same block of convolution layers as the first proposed model in Section 3.3.1. Flattened outputs of this block are input to a single gated recurrent unit layer with a hidden state of dimension 128. A fully connected layer projects the recurrent hidden state at time t to 3 output values. Again, a sigmoid activation is applied to these outputs and presence predictions are obtained at evaluation by applying a threshold value of 0.5. The model contains a total of about 3 million parameters.

3.3.3 Training procedure

The two proposed models, respectively the convolutional architecture in Section 3.3.1 and the recurrent architecture in Section 3.3.2, are trained on the development set described in Section 3.2. The dataset is split into training

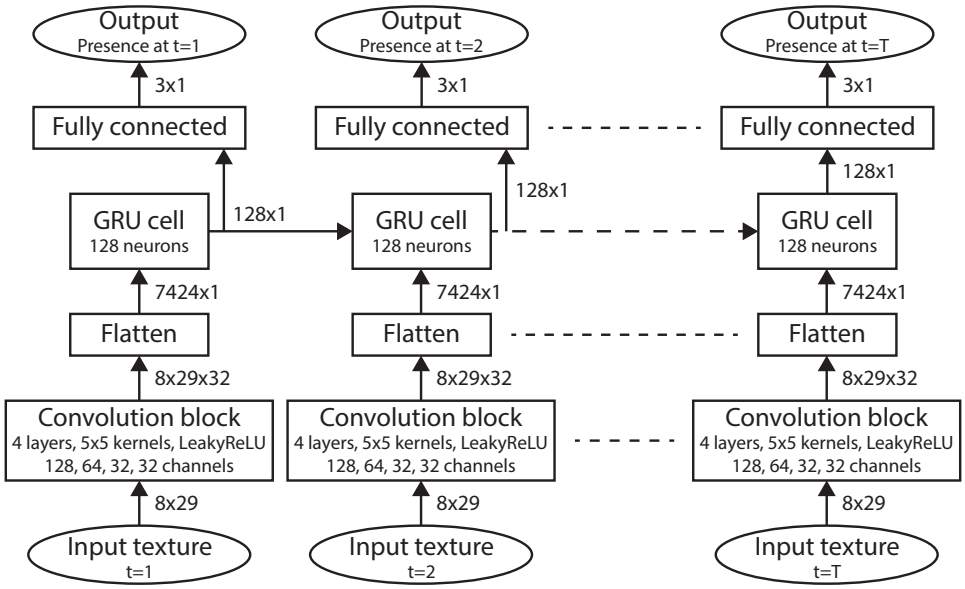


Figure 3.3: Architecture of the proposed recurrent neural network for source presence prediction in 1 s textures of third-octave fast measurements (unrolled view). Parameters are shared along all timesteps.

and validation subsets containing 70% and 30% of the 400 simulated scenes, thus 280 and 120 scenes respectively. For the proposed convolutional network an input example consists in a 8x29 texture corresponding to 1 s of third-octave fast measurements. To maximize the number of training examples, 1 s textures are extracted from simulated scenes with a hop (temporal step) size of 125 ms. This process is illustrated in Figure 3.4. The total number of 1 s textures in the training subset is thus about 10^5. This corresponds to the number of training examples for the convolutional network, whereas the recurrent network takes sequences of consecutive textures as input to learn dependancies over time. The length of sequences processed by the recurrent network is set to $T = 16$ in all experiments, as a result the recurrent network is trained on about 6000 example sequences of 2.875 s each.

Outputs of the models are compared to the binary labels for the presence of traffic, voices and, birds annotated by the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ indicator on corresponding 1 s textures. To do so, the loss function is the binary cross-entropy:

$$BCE(y, \hat{y}) = -\sum_s y_s log\left(\hat{y}_s\right) + (1 - y_s) log\left(1 - \hat{y}_s\right) \tag{3.6}$$

where s is the source, y_s and $\hat{y}_s$ are the target and predicted presence for

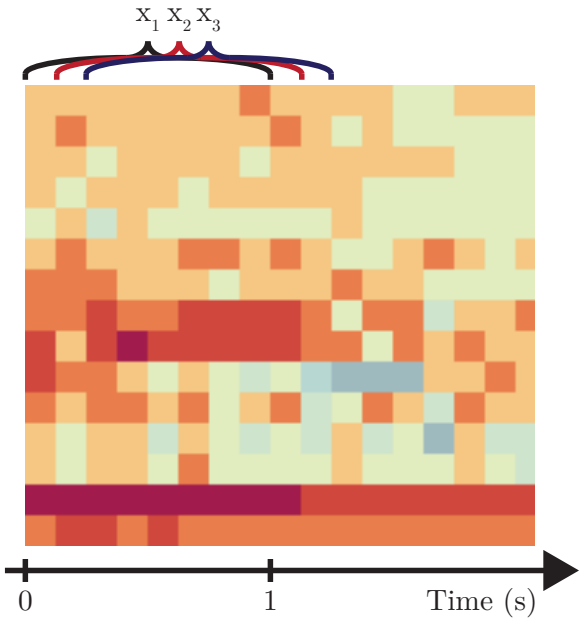


Figure 3.4: Extraction of 1 s textures x_n with 875 ms overlap from third-octave measurements, input sequentially to the recurrent neural network.

source s where y_s is binary and $\hat{y}_s$ is in the $]0-1[$ range.

The binary cross-entropy loss is minimized using the Adam algorithm [98], a popular stochastic gradient descent method. This algorithm iteratively applies small updates to model parameters from the gradient of the loss function as well as its first and second order moments, summarized over batches of examples. Default Adam hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are taken in all experiments, and parameter updates are applied with a learning rate $\lambda = 0.0001$. The batch size is 128 examples for the convolutional neural network, and 8 for the proposed model with recurrence to compensate the sequence length of 16. Thus, both models require the same number of training iterations to process the entire dataset.

Both models are trained for 20 epochs, *i.e.* iterations over the entire training dataset. This corresponds to about 15000 total iterations of the gradient descent algorithm. At the end of each epoch the loss is computed on the validation subset to ensure that the model has not overfitted on the training data. Figure 3.5 shows the evolution of the training and validation losses for both proposed models. At the end of the training process, the

training loss has converged for both models. The convolutional network (left) quickly overfits as the validation loss starts increasing from epoch 6 while the training loss still decreases. The state of the model at the end of epoch 5 is thus retained for evaluation. Similarly the model with recurrent decision (right) overfits, although only from epoch 13 to the end of the training, which can be explained by the higher complexity of this model. Parameters values are taken at the end of epoch 12 for evaluation. Validation losses for the two models have similar minimal values, respectively 0.148 and 0.131, and similar overall evaluation performances are thus expected.

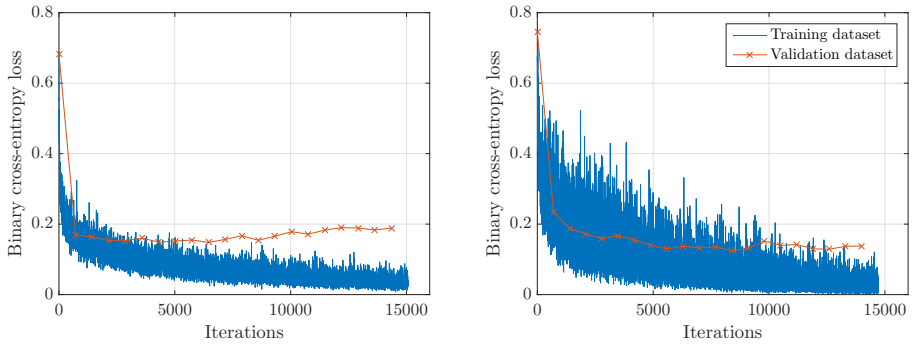


Figure 3.5: Evolution of training and validation losses of the convolutional (left) and recurrent (right) neural networks.

3.4 Evaluation

3.4.1 Source presence predictions

The performance of the two proposed models is first investigated on the evaluation dataset. The evaluation set is composed of 200 simulated scenes each containing 344 1 s textures overlapping by 875 ms. The convolutional neural network processes each texture separately, whereas the recurrent neural network processes a sound scene as a single sequence, thus its prediction at time t summarizes information propagated from all previous textures. Accuracy metrics are computed by comparing texture-specific predictions of source presence to corresponding $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ labels (n=68800). Table 3.1 summarizes the prediction accuracy metrics, including overall accuracy and true positive, true negative, false positive and false negative prediction rates. Both deep learning models perform well with over 90% accuracy on all

Table 3.1: Performance of the predictions of source presence by deep learning models trained with binary ground truth labels $\hat{T}_s(\alpha_{opt}, \beta_{opt})$. Presence metrics in % are computed for n=68800 1 s frames and time of presence metrics on n=200 45 s scenes. (TP: true positive, TN: true negative, FP: false positive, FN: false negative)

Convolutional network	All sources	Traffic	Voices	Birds
Presence accuracy (%)	91.87	92.62	93.34	89.64
Presence TP (%)	90.76	89.98	94.46	87.55
Presence TN (%)	93.24	98.98	92.11	91.23
Presence FP (%)	6.76	1.02	7.89	8.77
Presence FN (%)	9.24	10.02	5.54	12.45
$\hat{T}_s(\alpha_{opt}, \beta_{opt})$ RMSE	0.12	0.16	0.06	0.12
Recurrent network	All sources	Traffic	Voices	Birds
Presence accuracy (%)	93.67	96.15	92.96	91.90
Presence TP (%)	94.81	95.44	97.64	90.34
Presence TN (%)	92.26	97.86	87.84	93.09
Presence FP (%)	7.74	2.14	12.16	6.91
Presence FN (%)	5.19	4.56	2.36	9.66
$\hat{T}_s(\alpha_{opt}, \beta_{opt})$ RMSE	0.09	0.08	0.08	0.10

sources, indicating that their low complexity is still sufficient to perform the presence prediction task. The recurrent architecture improves accuracy by almost 2% (91.87% to 93.67%) at the cost of higher computation costs and number of parameters.

The recurrent network is more permissive on traffic and voice sources as shown by higher false positive rates (1.02% to 2.14% and 7.89% to 12.16%) and lower false negative rates (10.02% to 4.56% and 5.54% to 2.36%) for the two sources. The false negative rate of traffic detection is lower for the recurrent neural network than for the convolutional network, respectively at 4.56% and 10.02%. This indicates that including past context in the decision process with parametric recurrence helps identify traffic events that typically span several seconds. An increase in performance is also seen in bird presence prediction (89.64% to 91.90%), part of which may be due to the better identification of traffic events that sometimes contain high-frequency components mistaken for birds by the convolutional network. Still, the accuracy is lowest for bird presence detection in both models. An explanation is that the time and frequency resolutions of the input representation is too low for some patterns characterizing bird activity to appear.

Table 3.2: Pearson's correlation coefficients between the time of presence estimated by averaging presence labels predicted by the proposed deep learning model over time, and subjective annotations obtained during the listening test. (n=94)

Source presence indicator	Traffic	Voices	Birds
Convolutional network outputs	0.88**	0.87**	0.79**
Recurrent network outputs	0.84**	0.85**	0.83**
Ground truth $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ annotations	0.82**	0.82**	0.91**

**: $p<0.01$

$\hat{T}_s(\alpha_{opt}, \beta_{opt})$ time of presence estimates are obtained by averaging source presence predictions over 45 s in each simulated scene of the evaluation dataset (n=200). The overall root mean squared error (RMSE) compared to automatically annotated labels is higher with the convolutional network (0.12) than with the recurrent network (0.09) on a 0-1 time of presence scale. This is mainly due to the improved accuracy on traffic activity detection with the recurrent model, which leads to an RMSE of 0.08 compared to 0.16.

3.4.2 Application to the estimation of subjective attributes

Predictions of perceived source presence are applied to the estimation of the perceived time of presence of sources and pleasantness. To do so, predictions are compared to subjective assessments obtained on the listening test corpus in Section 2.2.3. The listening test corpus contains 6 recorded, 19 replicated and 75 simulated sound scenes which are evaluated separately to assess the generalisation capabilities of the trained models. Simulated sound scenes in this set are similar in construction and content to the deep learning development and evaluation datasets of Section 3.2. However, the simulation process involves a different database of isolated samples. The performance of deep learning models on this subcorpus should thus be similar to metrics computed on the evaluation corpus. Replicated sound scenes are also generated by *simScene*, but they are characterized by more complex scenarios and polyphonies, with additional sources not present in the training dataset (e.g. planes, construction work) annotated in [60]. Lastly, performance metrics computed on recorded sound scenes should provide an insight on whether the trained models can correctly identify sound sources in real-life sound environments.

Metrics of presence prediction accuracy cannot be computed on the lis-

tening test corpus as subjective assessments are on the time of presence over 45 s. Estimations of the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ indicator are therefore obtained by averaging source predictions output by the deep learning architectures over time. These estimations are compared to assessments of the perceived time of presence using the Pearson correlation coefficient. Table 3.2 shows results for the two proposed architectures compared to ground truth $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ values automatically annotated from separate source contributions. These labels are not available for the 6 recorded scenes, which are thus excluded from the analysis for fair comparison with deep learning model predictions. For the two proposed models, correlations between averaged presence predictions and the subjective time of presence is higher than for ground truth annotations on traffic and voices sources. Although this is not expected, a possible explanation is that the ground truth annotations do not correlate entirely with subjective assessments (r=0.82), and some wrongful predictions by the deep learning models are thus due to false labels. This would indicate that the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ indicator is too permissive on traffic sources leading to high false negative rates in learned model predictions, and similarly that it suppresses some voices activity leading to higher false positive rates in the predictions. However, the correlation is lower for both learned models on birds sources, which is expected as the corresponding accuracy is lower than for other sources. Because the correlation between the annotation and the subjective time of presence of birds is very high (r=0.91), bird activity misdetected by the model is likely to have a negative impact on the correlation. As previously noted in Section 3.4.1 the recurrent network has more consistent results across sources, and it performs slightly worse than the convolutional network on traffic and voices but better on bird activity.

Pleasantness predictions are then obtained by substituting the $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ estimations of relevant sound sources to the best model from acoustic variables found in Section 2.4.2, *i.e.* the $\hat{P}_{1,\varphi}$ in eq. 2.8. Thus, only predictions of birds presence are taken into account. Pleasantness estimations from outputs of the two proposed deep learning architectures are compared to the $\hat{P}_{1,\varphi}$ model from ground truth $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ annotations in Table 3.3. Predictions from both deep learning models perform comparably to ground truth source presence labels when applied to the pleasantness model. Interestingly, pleasantness estimations are better on the replicated subset for all three methods, despite the presence of additional sources that are not present in simulated scenes on which deep learning models are trained. The two deep learning architectures also yield higher pleasantness estimation errors on recorded scenes than on simulated scenes. This is expected as recorded scenes are more complex in terms of polyphonies and interactions between

Table 3.3: Quality of pleasantness predictions on the listening test corpus using deep learning models to predict source presence compared to ground truth 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	Sub-corpus	RMSE	r	R^2_{adj}
$P_{1,\varphi}$ with convolutional network outputs	All	0.87	0.90**	0.81
	Rec.	1.04	0.90**	0.50
	Rep.	0.70	0.92**	0.78
	Sim.	0.89	0.89**	0.77
$P_{1,\varphi}$ with recurrent network outputs	All	0.86	0.90**	0.81
	Rec.	1.05	0.92**	0.50
	Rep.	0.68	0.93**	0.79
	Sim.	0.88	0.89**	0.78
$P_{1,\varphi}$ with ground truth $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ labels	All	0.83	0.91**	0.82
	Rep.	0.72	0.92**	0.79
	Sim.	0.86	0.89**	0.79

**: $p<0.01$

sources. Still, the pleasantness root mean squared error on this sub-corpus is about 1.05 for the two models, which is below the average standard deviation in individual assessments of 1.77 (see Section 2.4.2 for details). No direct comparison with $\hat{T}_s(\alpha_{opt}, \beta_{opt})$ labels is available on recordings and metrics are computed on n=6 sound scenes only. While these results are promising, evaluation on a larger corpus of subjectively annotated recorded scenes would thus be necessary to conclude on the generalisation capabilities of the trained models.

3.4.3 Application to sensor data in Lorient

The proposed deep learning models are applied to sensor data from the Cense network in Lorient. Specifically, the time of presence of traffic, voices and bird sources is estimated on the 20th of August, 2020 between 22h and 22h10. Data from 37 sensors active during this time period is dynamically requested from storage servers and processed by the deep learning architectures on a remote computer. Figure 3.6 shows maps of the time of presence estimations at sensor locations by the convolutional (top) and recurrent (bottom) neural networks. Including data querying, the total computation time was about

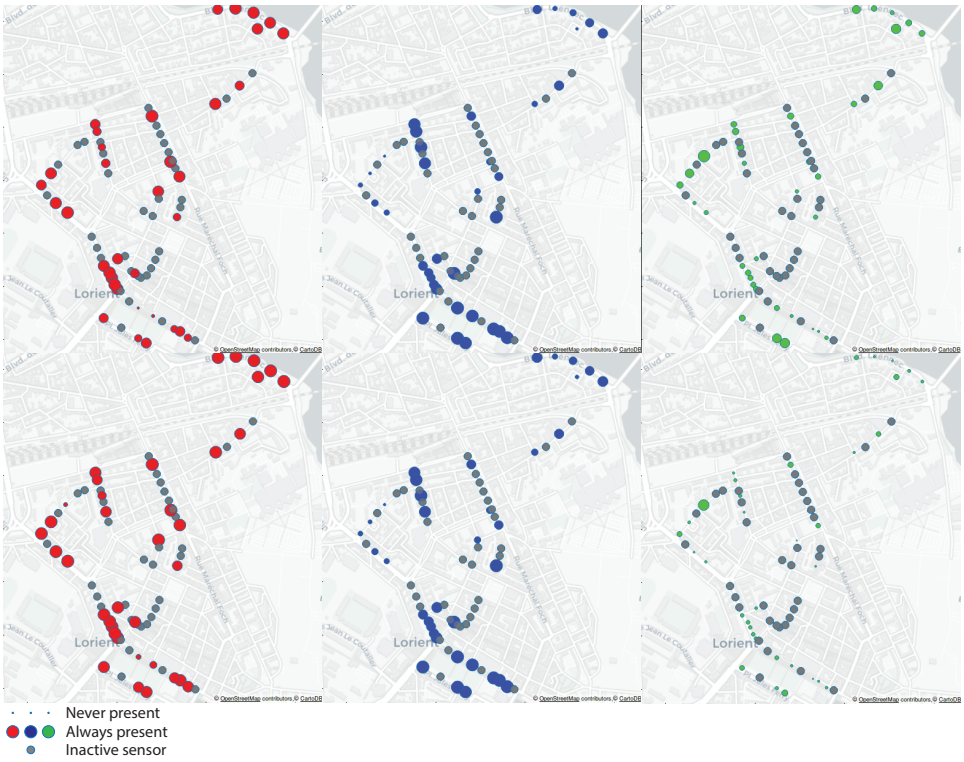


Figure 3.6: Maps of predictions by the convolutional (top) and recurrent (bottom) networks of the time of presence of traffic (left), voices (middle), and birds (right), for data collected by the Cense network on the 20th of August 2020 between 22h and 22h10.

1 min for the convolutional network and 3 min for the recurrent network. Because of the data query format each sensor was processed sequentially. Running time of presence estimation models on data servers would allow fully parallelizing the processing of measurements from multiple sensors, further reducing the computation durations significantly. The proposed approach can thus be regarded as scalable in the context of continuous monitoring with large-scale sensor networks.

The two architectures behave similarly on sensor data and simulated scenes of the evaluation dataset in Section 3.4.3, as the recurrent network predicts more traffic and voices activity and less bird activity compared to the convolutional network. Overall, the predictions correspond to expectations, for example high voice activity on Franchet d'Esperey Bvd. (north-

west on the map) and Jules Ferry Sq. (south) with several restaurants and cafes, traffic on the main streets and bird activity matching green spaces locations. However, quantitatively evaluating the quality of predictions is difficult because sensors do not record waveform audio. Thus, no perceptual assessments are available on sensor data. Recording and annotating acoustic scenes in Lorient is necessary to fully evaluate the generalisation of trained models.

In general, the performance of deep learning models is limited by the quantity and quality of available training data. First, the quantity of training data refers to the amount of unique scenarios and spectral patterns in the dataset. The quantity of scenarios that can be generated is theoretically infinite, although redundancy is introduced by the relatively low size of the isolated samples database, on which the diversity of spectral patterns in the training dataset depends. Second, the quality of training data refers to both its diversity and its closeness to sound sources and environments encountered in the application. The scope of generated scenarios in simulated scenes is validated in Chapter 2 for a district of Paris and are not expected to differ significantly in the case of Lorient. However, the extracts composing the isolated samples database for simulating sound scenes may not match all sources of interest encountered in the studied environments. For example, seagulls are an important and near omnipresent sound source in Lorient, but are not represented in the isolated samples database. Wether seagulls should be perceptually categorized as birds is debatable and lies beyond the scope of the present study. In any case, they are characterized by unique spectral signatures which the deep learning models should be trained to recognize. Furthermore, time-frequency structures characterizing each source may vary in recordings made with different microphones due to their respective impulse responses. Here, the samples are gathered from FreeSound contributions and microphone types are unknown. Thus, ideally the isolated samples database would be composed of recordings made in Lorient, with the same microphones as in acoustic sensors.

Chapter conclusion

The developed deep learning models are able to predict source presence from sensor data, with high accuracy and faster than real-time speeds. Presence predictions yield low errors on the estimation of the perceived time of presence of sources of interest and pleasantness. However, these results are obtained by comparison to subjective assessments on a corpus mostly composed of simulated sound scenes. Simulated training data are constructed

from a location-independent isolated samples database, where some sources present in the target sound environments of Lorient are not represented. Chapter 4 thus investigates methods to improve and evaluate the adaptability of deep learning models, where knowledge about the specific application environments is introduced.

Chapter 4

Domain adaptation techniques for robust learning of acoustic predictors

In order to achieve the domain specialization of deep learning architectures trained on simulated data to target environments, a new set of experiments is conducted that investigates the relevance of transfer learning techniques to improve the performance of time of presence predictors.
To do so, latent audio representations are learned on a pretext task from large amounts of sensor data. This type of data has the advantage of corresponding to the application domain, but cannot be labeled for the target task.
A low-complexity recurrent decision architecture is then trained on simulated data to predict source presence from these robust latent representations. To do so, a second simulated training dataset is constructed from source-specific extracts recorded in sound environments of the application.
Three pretext tasks are investigated, including two discriminative supervised tasks and a regressive unsupervised task. Evaluation conducted on manually annotated on-site recordings demonstrates that audio representations learned on pretext tasks improve presence prediction accuracy when the domain correspondence, dataset duration or sample diversity of simulated corpora are limited.

4.1 Introduction

4.1.1 Design of deep learning architectures

On synthetic evaluation data, the deep learning architectures presented in Chapter 3 achieve excellent source presence detection accuracy, and predicted labels result in estimations of the perceived time of presence with low errors. However, the performance metrics are mainly evaluated on simulated acoustic scenes. These scenes are similar in distribution to sound scenes on which model parameters are optimized, but different from measurements collected by sensors in the final application. Here, the distribution (or domain) of data refers to its contents in terms of the sound source taxonomy, the covered spectral signatures representing sound sources as well as the range of possible polyphonies and scenarios. In particular, the taxonomy of sources in simulated scenes is incomplete. Deep learning models are thus not trained to identify spectral patterns associated to missing classes (e.g. construction works, water sounds) or some subclasses of sources of interest (e.g. bird species). Due to these limitations of simulated scenes, there is no guarantee that trained architectures are capable to generalize to real-life situations encountered in Lorient sound environments. Generalization properties of deep learning models depend on two necessary conditions. First, sufficient amounts of training data must be available, where quantity is defined by both the number of training examples and their diversity. Second, this data must belong to the same domain (distribution) as the data processed by trained models in the application. Because learning model parameters directly on Lorient sensor data is not possible, transfer learning techniques are instead investigated in this chapter, in order to evaluate their ability to localize detection tools based on deep learning techniques.

Transfer learning techniques exploit the design of most architectures aimed at solving discriminative tasks, which follows the principles depicted in Figure 4.1. An encoder part first processes the input signal and returns a latent representation. Its role is to extract and decorrelate useful information in the input representation, leading to a lower-dimensional representation of the signal. In addition, information that is not necessary to solve the task is ideally discarded or reduced. The latent representation should then allow a decision architecture with low complexity, *i.e.* low number of parameters and nonlinearities, to efficiently complete the task. Encoder-decoder models are similarly designed to learn a low-dimensional code containing sufficient information to reconstruct a high-dimensional signal. In these architectures the decision architecture is replaced by a decoder architecture with complexity

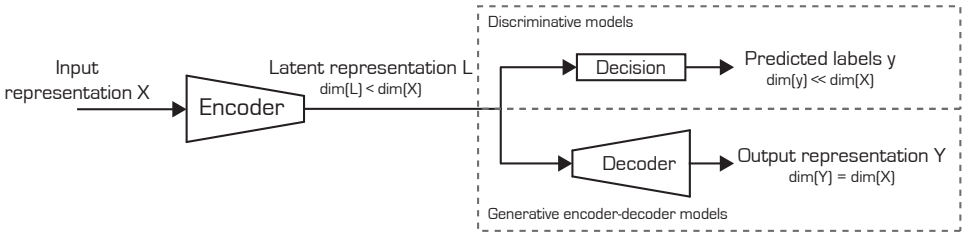


Figure 4.1: Typical architecture design of deep discriminative or encoder-decoder approaches. Similar encoder architectures are relevant to solve a variety of related tasks, whereas the decision or decoder is task-specific.

equivalent to that of the encoder.

Although its parameters are adapted to the task during the training process, the encoder architecture in discriminative or encoder-decoder approaches is task-agnostic, and typically composed of stacked convolutional layers corresponding to a highly nonlinear filterbank applied to the input signal. In contrast, the decision architecture is entirely motivated by the studied task. In discriminative models, the encoder comprises the majority of the final architecture's parameters, whereas the decision is composed of a limited number of fully connected, convolutional or recurrent hidden layers. An example is the source presence prediction architectures proposed in Chapter 3. Transfer learning and self-supervised learning approaches specifically utilize the general-purpose properties of the encoder.

4.1.2 Transfer learning

Transfer learning and self-supervised learning methods aim at solving problems in which only a limited number of examples are labelled for the desired task. To do so, these methods rely on the availability of large amounts of related data, either annotated for another task or unlabeled respectively. The objective is to learn robust latent data representations by extracting underlying information in large datasets. Optimal parameters of an encoder architecture trained to produce this representation are then a good initial state for parameters of the model solving the target task. In contrast, training a model from scratch (random initialization) on insufficient amounts of data compared to the network complexity would yield direct overfitting to training training examples.

Figure 4.2 summarizes the framework of transfer learning and self-supervised learning approaches. First, a model is trained to solve a pretext task on data

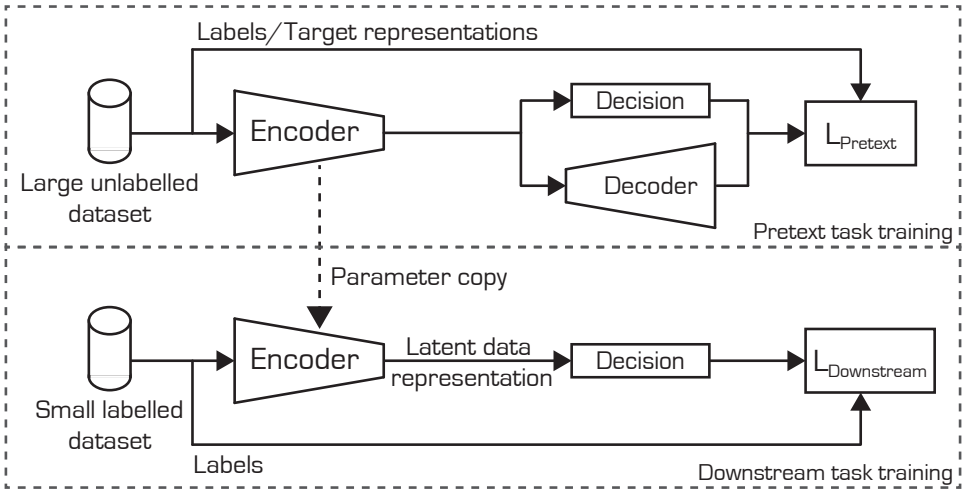


Figure 4.2: General framework of transfer learning approaches. An encoder is trained to solve a pretext task on a large dataset, and the same encoder architecture in the downstream task is initialized with obtained optimal parameters architecture, in order to stabilize training with few labeled examples.

unlabeled for the target task. Once the training process is completed, optimal parameters of the encoder are retained to initialize the same encoder architecture in the downstream (target) task model. Two approaches are then possible when training the model on the downstream task: i) freeze the pre-trained encoder parameters and learn only the decision sub-network, and ii) learn parameters of the entire model, *i.e.* fine-tune parameters of the encoder. In the first approach, the decision part has a low capacity due to a limited number of parameters and nonlinearities (depth). Thus, its efficiency depends on the closeness between the pretext and downstream tasks in terms of the needed information extracted by the encoder. If the two tasks are close, latent representations of the input signal should be similar for the corresponding optimal models. Conversely, if the two tasks are unrelated training a low-complexity decision sub-network may be insufficient to solve the downstream task. Fine-tuning encoder parameters allows us the latent representation to change in information content. In this case, the entire model, trained with a low amount of data with respect to the complexity of the network, overfits rapidly. However, because most of the model's parameters are initialized with an already good solution, the aim of fine-tuning the encoder is to achieve a better state before overfitting compared to starting

the training process with random parameters. Generally, the fine-tuning approach yields better performance. Still, we found that freezing the encoder weights may be preferable in the case where there are very few annotated examples available for the downstream task.

Domain adaptation refers to the subset of problems where the pretext and downstream tasks are the same, but available large datasets are not in the same domain as the target data (e.g. pitch classification with a large piano dataset adapted to a small violin dataset, or birdsong detection in two environments with different bird species). In that case, pre-trained parameters of the decision architecture may also provide a relevant initial state in the downstream task.

Pretext tasks can be formulated as supervised discriminative tasks (transfer learning) or unsupervised tasks associated with encoder-decoder architectures (self-supervised learning), depending on the availability of auxiliary labels in large datasets. Different information content is then found in the corresponding latent representation. The encoder in a discriminative model only retains information useful to solving the task. Thus, choosing related supervised pretext tasks is important to ensure that all necessary information for the downstream task is contained in the latent representation. In an encoder-decoder architecture, the latent representation is an information bottleneck as input and output representations have similarly higher dimensions. The main role of the encoder is thus to perform efficient dimensionality reduction of input signals. Latent representations in encoder-decoder models contain sufficient information to allow the estimation of high-dimensional signals, thus their information content should be sufficient to solve the downstream discriminative task. However, because the decoder architecture is typically as complex in terms of depth and number of parameters as the encoder, a low-complexity decision architecture may not be sufficient to link information in the latent representation to the desired outputs in the downstream task.

4.1.3 Pretext tasks for audio

Several pretext tasks have been investigated for unsupervised audio representation learning. In [16], the unsupervised learning of embeddings from sensor data is specifically addressed. The proposed method (*TriCycle*) relies on available timestamps metadata associated to sensor measurements as labels. A convolutional architecture first encodes the input audio into a latent representation. The decision architecture is an ensemble of three identical fully connected networks, that process the obtained embedding in addition

to sensor identifier information to respectively predict the time of day, the day of the week and the month of the year. Directly predicting scalar values for the three temporal cycles introduces issues of proximity (e.g. around midnight for the time of day). Instead, each position in temporal cycles is coded as a phase, and the model predicts the corresponding cosine and sine values. The obtained audio embedding improves the performance of models predicting the sensor identifier from audio as a downstream task.

When available, multimodal data can provide useful additional information in learned latent representations. In [99], the authors address the task of audio-visual correspondence in a dataset of videos. Examples are composed of an image and a sound extract extracted from the same or different videos at random. Separate convolutional encoder architectures extract information from the two inputs into embeddings, and a low-complexity fusion architecture predicts whether the inputs are matched. The encoder architecture processing audio content is then applied to train a support vector machine (SVM) on several classification tasks. The proposed approach achieves state-of-the-art performances on the environmental sound classification (ESC-50) dataset, as well as on ambiance classification with the DCASE2013 task 1 dataset. Design considerations for the multimodal approach are further investigated in [100]. The authors find that for environmental classification downstream tasks, the pretext task can be trained on data from a different domain (e.g. music). The most important factors on which the performance of learned latent representations depends are instead found to be the quantity of pretext training data, as well as the choice of input audio features.

Training an encoder architecture with a triplet loss is investigated in [101]. Triplets are formed by sampling two audio texture frames from the same example as well as a third texture frame from a different example. Each texture is encoded independently, and the training process minimizes the distance between latent representations of matching textures while maximizing the distance between embeddings of non-matching textures. Matching textures can also be obtained by transforming a reference sample, for example adding noise or time-frequency shifting.

An unsupervised regression pretext task is proposed in [102], that is inspired by the success of the *Word2Vec* representation learning method in natural language processing applications [103]. The approach is characterized by an encoder-decoder model where both parts are composed of recurrent layers. In the first setting of skip-gram, an audio texture is input to the model, which predicts an arbitrary number of plausible context textures, *i.e.* the N past and N future texture frames. In the second setting of continuous bag of words (CBoW) [104], inputs and targets of the model are swapped.

The context textures frames are encoded and the decoder produces an estimation of the central texture frame. The model, initially proposed for representation learning of speech signals, is evaluated against other methods on environmental data in [105]. Results show that when training the downstream task on a very low number of labeled examples (1000), the *Audio2Vec* approach outperforms discriminative pretext tasks with auxiliary data as labels, as well as the triplet loss.

Lastly, the authors of [106] develop an encoder architecture simultaneously trained on multiple pretext tasks. An input speech segment in the waveform domain is encoded by a convolutional architecture, and the resulting embedding is input to several fully connected architectures each associated to a task. Some of these tasks are regressive with common handcrafted features as targets, including the signal waveform, Mel frequency cepstral coefficients and the log-power spectrum. Other tasks are discriminative, such as determining whether two examples are produced by the same speaker. This combination of simultaneously trained tasks yields well-generalized and task-agnostic latent representations of the audio signal.

4.1.4 Localized embeddings

In the current study, large amounts of third-octave data are available from the Cense sensor network in Lorient. However, corresponding annotations for the target task of source presence prediction are not available. Relevant information about the target domain of sensor data can thus only be inferred by learning latent audio representations on pretext tasks, and the presence prediction task can then be trained on simulated data, for which labels are computed automatically. Acoustic scenes simulated using the *simScene* process are incomplete in terms of the scenarios, polyphonies, and diversity of spectral signatures associated to sound sources in isolated samples. This setting is different from typical transfer learning and domain adaptation applications, in which the downstream task is always trained on data from the target domain.

In this study, we hypothesize that transfer learning techniques enables the localization of models trained on simulated datasets. Pre-training an encoder on large amounts of sensor data yields a latent audio representation well-generalized to the target domain of Lorient sound environments. Despite part of this information being lost when training the presence prediction task on simulated sound scenes, knowledge about Lorient-specific polyphonies and spectral patterns should remain in the final source presence prediction architecture with optimal parameters. In this regard the proposed approach

is still akin to domain adaptation.

However, several considerations must be observed. First, the implications of fine-tuning or freezing the encoder parameters learned on pretext tasks are not as straightforward as in typical transfer learning approaches. On one hand, because the downstream task is not trained on data from the target domain, fine-tuning encoder parameters trained to optimality on sensor data may be detrimental, because the encoder forgets information not contained in simulated data. On the other hand, the latent representation is nearly optimal for solving the pretext task, but is certainly suboptimal to solve the downstream task. Freezing encoder weights increases the difficulty of solving the presence prediction task with a low-capacity decision architecture. In any case, it is important that the distribution of simulated data is as close as possible to that of sensor data. If so, fine-tuning the encoder parameters while learning the downstream task leads to only a small amount of the target distribution, learned during pretext task training, being lost. In practice, the diversity of environments covered by simulated datasets can be artificially increased by simulating arbitrarily more scenarios. Arbitrarily more labeled data are thus available to train the downstream task compared to typical transfer learning. However, the correspondence between simulated and recorded data distributions is limited by the quality of the isolated samples database and the scenario generation parameters in the simulation process.

Furthermore, the pretext and downstream tasks should be as close as possible in terms of the necessary information to solve them. Two closely related tasks will likely yield similar information content in optimal latent representations of the audio data. Thus, if the information needed to solve the pretext and downstream tasks is sufficiently close, the learning process operates fewer modifications of the encoder to obtain an optimal latent representation for the the downstream task. Pretext tasks that require the propagation of more information than the downstream task, such as generative tasks, are an interesting alternative. The role of the decision architecture in the downstream task then becomes to filter useful properties from very informative learned latent representations of audio. In this case, fine-tuning the encoder parameters is less important as the latent representation likely contains enough information to solve the presence prediction task, and freezing them to retain knowledge generalized to Lorient sensor data may be preferable.

Figure 4.3: Map of sensors implemented as part of the Cense network in Lorient. The *SpecCense* dataset is composed of data from the 16 sensors shown in blue.

4.2 Audio representation learning from sensor data

4.2.1 Large dataset of sensor measurements

A large dataset is constructed for the purpose of training pretext tasks. This dataset, termed *SpecCense* in the following sections, is composed of measurements from the Cense sensor network in Lorient. Each sensor records the timestamped third-octave spectrum in the $[20Hz - 20kHz]$ range every 125 ms. This dataset was constructed from the first available measurements following the implementation of the sensor network, with technical issues and power failures leading to high temporal discontinuity in stored data. 16 sensors among 74 in the network were automatically selected that maximized the temporal continuity of available spectral data. The location of all sensors is shown in Figure 4.3, and selected sensors are highlighted as blue points. In practice, sensors on the same street are connected to the same controller which sometimes failed in the early network implementation. Sensors selected to maximize measurement continuity thus correspond to specific controllers, and only cover part of the target sound environments. In future

work, deep learning datasets should instead be constructed by eliminating information redundancy yielded by sensor proximity, and maximizing the diversity of sound environments and polyphonies.

The *SpecCense* dataset initially comprises third-octave measurements collected by the 16 selected sensors between December 1st, 2019 and December 5th, 2019. In theory, this results in about 2000 h of audio data. However, to facilitate data processing in the training of deep learning models, a preprocessing step is applied that eliminates non-continuous blocks of measurements. Due to various transient failures of the network, timestamps of adjacent third-octave frames may be separated by more than 125 ms. Here, a tolerance of 125 ms is allowed for continuity, *i.e.* two adjacent frames for which the timestamps are less than 250 ms apart are regarded as continuous measurements. Even within the subset of selected sensors, longer downtime periods occur where no measurements are available, due to power malfunction or on-board computation bottlenecks. Thus, the final dataset contains a total of 1280h of spectral data. This duration is still two orders of magnitude higher than datasets typically considered for event detection, and 25 times longer than the dataset of simulated scenes on which the presence prediction model in Chapter 3 is trained.

The dataset is split into training, validation, and evaluation independent subsets to respectively perform optimization, monitor generalization, and evaluate the performance of models on pretext tasks. The split is done along time: the training subset contains the first 70% of measurements for each sensor, the validation and evaluation subsets contain the next 10% the last 20% of measurements respectively. Note that because measurements are not available across the entire measurement period for some sensors, the timestamps separating data from different subsets may differ for each sensor.

4.2.2 Encoder architecture

In Section 4.1.4, the models solving the pretext and downstream tasks share an identical encoder architecture, and encoder parameters in the downstream task are initialized with optimal values obtained by training on the pretext task. The same paradigm is adopted in the current study, where the downstream task is the prediction of source presence from spectral sensor measurements. The proposed encoder architecture takes as input a texture frame of 1 s of fast third-octave measurements (8×29) and outputs a real-valued vector (128×1) as a latent representation of the audio data.

The architecture of the proposed encoder is shown in Figure 4.4. It is composed of 6 convolutional layers characterized by 3×3 filters and 64, 64,

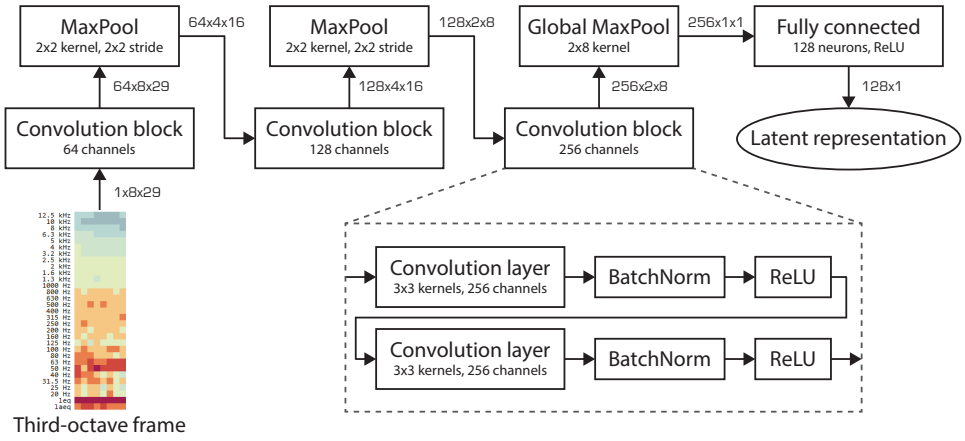


Figure 4.4: Proposed encoder architecture to extract information from 1 s third-octave texture frames. This architecture is common to all studied models in a transfer learning setting.

128, 128, 256 and 256 separate output channels respectively. Convolutional layers are followed by batch normalization layers [107] that normalize each batch of input examples, then apply learned scale and shift parameters to the result:

$$y = \gamma \left(\frac{x - \mu_x}{\sqrt{\sigma_x^2 + \varepsilon}} \right) + \beta \tag{4.1}$$

where x and y are the input and outputs of the layer, γ and β are learned scale and shift parameters, μ_x and σ_x are the mean and standard deviation of x computed along the examples dimension in the batch and ε is a small constant for numerical stability. Batch normalization layers stabilize the training process by removing the potential variance in level observed in different batches of examples. This typically helps model parameters converge with fewer iterations, which is desirable when training architectures on very large datasets. A rectified linear unit (ReLU) element-wise activation is further applied to the output of batch normalization layers to introduce nonlinearity to the model. Groups of two convolutional layers each followed by batch normalization and ReLU layers are designated convolution blocks. After each convolution block the representation is downsampled by a factor of two in the time and frequency dimensions, by applying a max-pooling operator to each channel independently. For all 2×2 non-overlapping groups of adjacent values in the representation, max-pooling returns the maximum value of this group. This is an alternative to strided convolution layers pre-

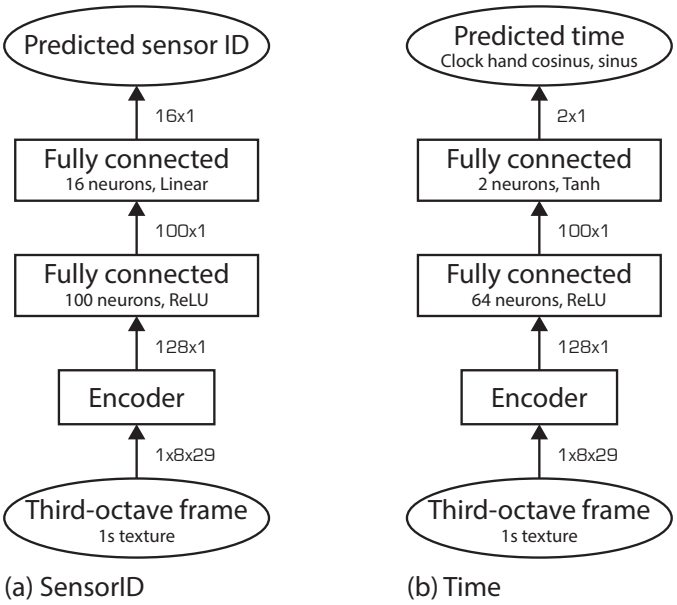


Figure 4.5: Proposed architectures for the *SensorID* and *Time* pretext tasks models.

sented in Section 3.3, in which filters are moved by more than one step during convolution to downsample data representations. The max-pooling operator following the last convolution block reduces the time-frequency representation in each channel to a scalar value. As a result its output is a vector in $\mathbb{R}^{256}$, where 256 is the number of output channels in the last convolutional layer. Lastly, a fully connected layer transforms this vector into another with the desired embedding size, which is set to 128 in this study. This output is the latent representation that is input to task-specific decision or decoder architectures. The total number of parameters in the encoder architecture is about 1.2 million.

4.2.3 Supervised pretext tasks

Two supervised pretext tasks are first investigated to train the encoder architecture on the *SpecCense* dataset, in order to obtain latent representations of audio data that are well generalized to Lorient environments. These tasks are discriminative and rely on metadata available with sensor measurements. Specifically, the use of sensor identifier and timestamp informations is proposed.

The first pretext task under study consists in predicting the sensor identifier associated to an input 1 s third-octave texture frame. This is a single-label classification task with 16 classes, and is termed *SensorID* in this chapter. The architecture of the proposed deep learning model for this task is shown in Figure 4.5(a). Input 1 s third-octave texture frames are first encoded by the architecture presented in Section 4.2.2. The resulting latent representation is processed by a low-complexity decision architecture that predicts a real-valued vector of dimension 16. The decision part is composed of two fully connected layers. The first layer is followed by a rectified linear unit nonlinear activation applied element-wise to the data. The second layer has no activation function, and its output is instead passed through a softmax layer that scales the predicted values so that their sum equals 1. The scaled vector can then be interpreted as a discrete probability distribution where each value corresponds to a class. The network is trained with a cross-entropy loss to minimize the negative log-likelihood of predictions. If i is the active ground truth label, *i.e.* the correct class, this yields:

$$\hat{y}_i = softmax(x_i) = \frac{e^{x_i}}{\sum_{i=1}^{I} e^{x_i}} \tag{4.2}$$

$$CE(\hat{y}_i) = -log(\hat{y}_i) \tag{4.3}$$

where I is the number of classes, and x_i is the output of the last fully connected layer for class i.

The second pretext task consists in estimating of the time of the day at which a given 1 s third-octave texture frame is recorded. This task is referred to as *Time* in this chapter. The problem is cast to the prediction of two values in the $[-1, 1]$ range in the TriCycle model [16], which corresponds to the cosinus and sinus of the phase of the hour clock hand. Compared to a scalar prediction of the time, this solves the discontinuity problem that occurs at midnight. The proposed architecture is shown in Figure 4.5(b). It is very similar to that of the *SensorID* task, although the first fully connected layer has a slightly lower number of neurons (64) than for the *SensorID* task (100) to account for the lower number of output predictions. Outputs of the last layer are compressed into the $[-1, 1]$ range by applying an element-wise hyperbolic tangent (Tanh) activation. Network parameters are then optimized to minimize a mean squared error loss function comparing predicted values to the sinus and cosinus of the ground truth time of day, as given by timestamps.

Both models are trained on the *SpecCense* dataset using the Adam algorithm with default parameters and a learning rate of 0.0001. The models

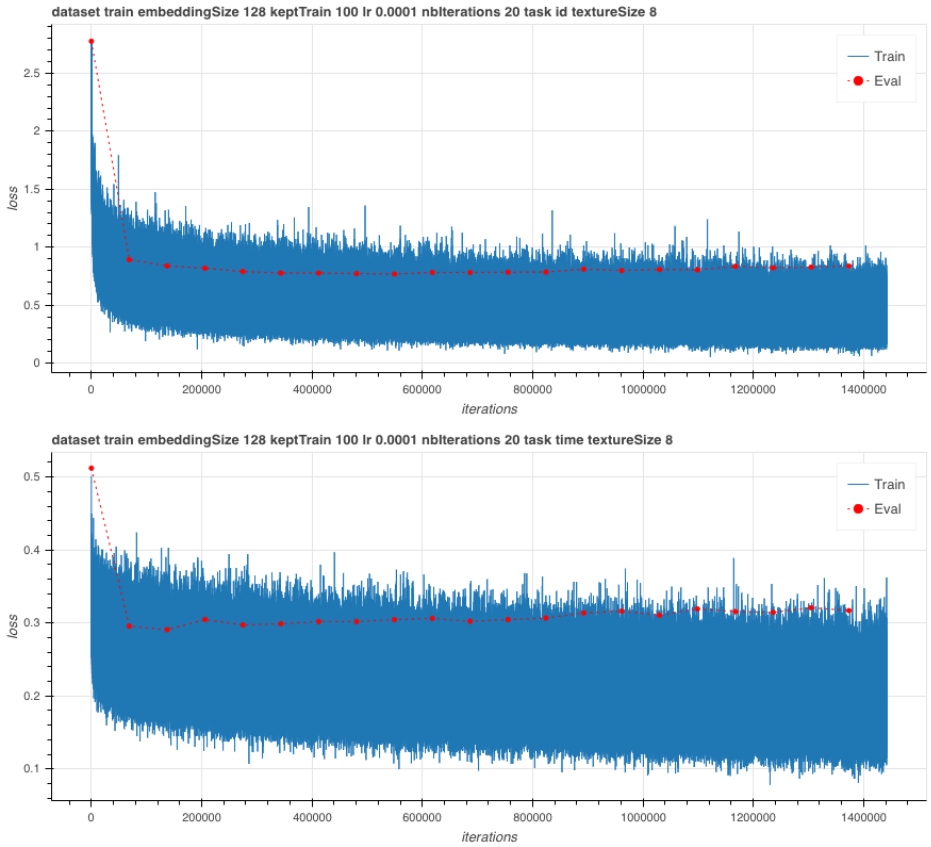


Figure 4.6: Evolution of training and validation losses for the *SensorID* and *Time* pretext task models.

are trained for 20 epochs with batch size 64, which corresponds to about 140000 iterations. Figure 4.6 shows the evolution of training and validation losses for the *SensorID* (top) and *Time* (bottom) tasks. Both models rapidly converge, then slowly overfit to the training data. Model states yielding the lowest validation losses are retained, at the end of epoch 8 for *SensorID* and epoch 2 for *Time*. On the evaluation dataset, these models yield about 75% sensor identification accuracy and 3 h time estimation error respectively.

4.2.4 Unsupervised pretext task

In addition to the two discriminative pretext tasks, the *Audio2Vec* task is considered as a generative alternative [105], which does not require the avail-

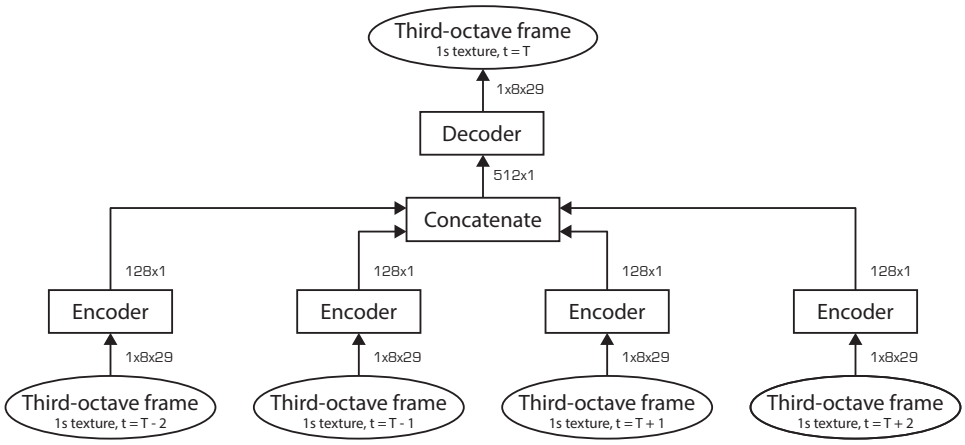


Figure 4.7: Encoder-decoder architecture proposed to solve the *Audio2Vec* pretext task by predicting a third-octave texture frame from context texture frames.

ability of auxiliary labels. In discriminative models the encoder extracts information useful to the prediction into a low-dimensional latent representation, and discards information content that does not contribute to solving the task. The objective of the encoder in a generative encoder-decoder model is different, as it must retain enough information to reconstruct a signal in the same domain and dimensions as its inputs. Thus, in an encoder-decoder architecture the latent representation is an information bottleneck, and the encoder's focus is more oriented towards dimensionality reduction than information filtering. Studying the effect of these information content differences in the context of transfer learning is of interest to the current study.

The *Audio2Vec* task consists in predicting the 1 s third-octave texture frame at time t given the previous P and next P texture frames. This is illustrated in Figure 4.7 for a context size $P = 2$. The input and output texture frames are continuous but do not overlap. This task is entirely unsupervised, as it does not rely on metadata recorded by sensors. The associated architecture is an encoder-decoder model, with each context texture frame processed independently by the encoder, yielding $2P$ separate latent representations. The concatenation of these representations are input to a decoder with similar complexity to the encoder, both in terms of parameter number and nonlinearity level.

The architecture of the decoder is shown in Figure 4.8. It mirrors approximately that of the encoder. A fully connected layer first transforms the

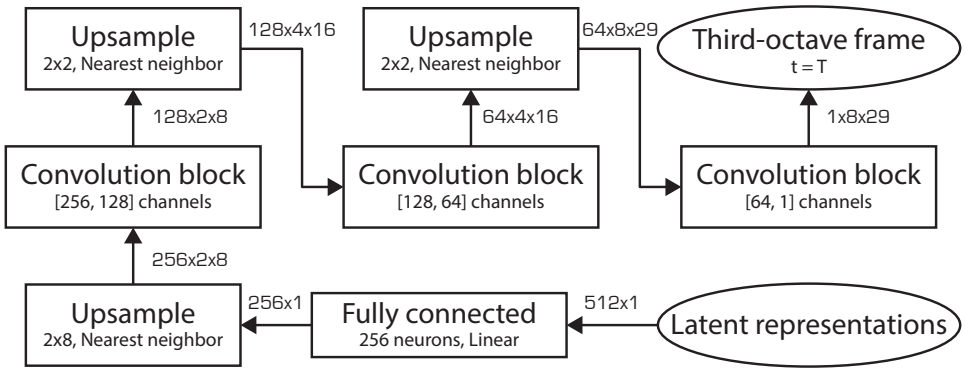


Figure 4.8: Decoder architecture proposed to solve the *Audio2Vec* pretext task.

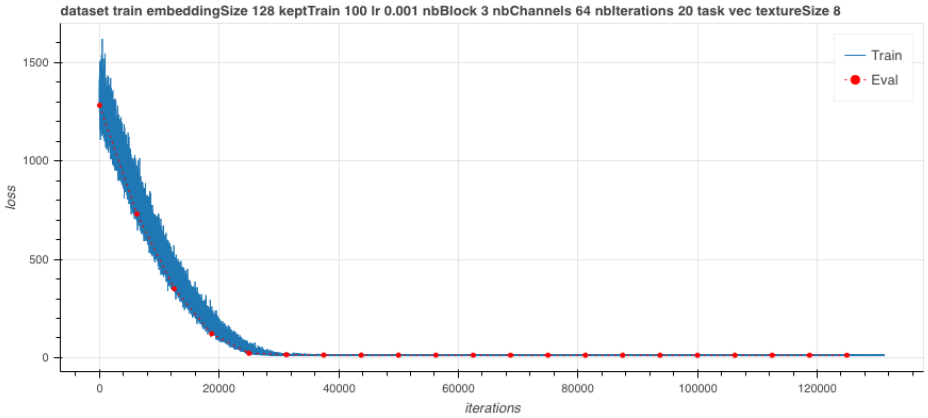


Figure 4.9: Evolution of the training and validation losses of the *Audio2Vec* model trained on sensor data.

concatenated latent representations to a vector of size 128. Max-pooling layers are replaced by nearest-neighbor upsampling layers with the same kernel size to revert their effect in terms of dimension changes. Convolution blocks are composed of two convolution layers with 3x3 kernels followed by batch normalization layers and rectified linear unit activations, with reversed number of channels (resp. 256, 128, 128, 64, 64, 1). The model is trained using a mean squared error loss function comparing the predicted and ground truth third-octave texture frame on each time-frequency cell.

Model parameters are optimized on the *SpecCense* dataset using the Adam algorithm with default parameters and a learning rate of 0.001. This

value is higher than in the training process of previous models in order to accelerate convergence, as the process was stable but slow with a learning rate of 0.0001. Optimization is performed for 20 epochs with batches of 64 examples. In discriminative pretext tasks, an example is available for each 1 s third-octave texture frame in the dataset. To obtain a similar number of iterations, each 1 s texture frame appears as the target in a training example. Thus, a given texture frame is part of multiple examples as either an input or the target. As shown in Figure 4.9, the training and validation losses converge to very low errors after a few epochs. Interestingly, the model does not appear to overfit to training data. Thus, optimal parameters are taken at the end of the training process.

4.3 Presence prediction with transfer learning

4.3.1 Local controlled dataset

The downstream task of source presence prediction on 1 s third-octave texture frames is investigated, using the encoder architecture pre-trained in Section 4.2 on pretext tasks. Ideally, deep learning models for this task would be trained on measurements from the Cense sensor network. However, annotations for this task are not available on sensor data. Thus, the models are trained on sound scenes simulated with *simScene*, for which annotations can be computed. In Chapter 3 a dataset is simulated for this purpose from isolated samples associated to sources of interest. These isolated samples are obtained from the Freesound database, or from other publicly available datasets. The corresponding recordings are made with a wide of microphones whose characteristics are largely unknown, and does not necessarily matches sources found in the application environment of Lorient (e.g. some bird species or vehicle types). Thus, this dataset is referred to as *TVBUniversal* in the remainder of this chapter. From the discussion in Section 4.1.4, the dataset on which the downstream task models are trained should be as close as possible to sensor data recorded in Lorient in terms of distributions. When simulated sound scenes, the two major aspects that can be studied to improve this correspondence are i) the curation of the database of isolated samples and ii) the careful setting of the design parameters for the generation of scene scenarios. To address the first case, a database of isolated samples specific to Lorient environments and sources is constructed from short on-site recordings. Adapting scenario generation parameters to Lorient requires the annotation of source activity for a large number of acoustic scenes for several ambiances to obtain relevant distributions, similarly to Section 2.2.2.

This process is not feasible on sensor measurements as no waveform audio is recorded. Although longer on-site recordings could be obtained and annotated manually, this study does not address this approach due to the time consumption involved. Thus, this dataset termed *TVBCense* contains scenes with temporal distribution of events typical of urban areas, but not specifically tailored to Lorient. In [108], this was shown to have a lower impact on generalization with respect to the isolated sound dataset.

To build the isolated sound dataset, a set of 10 sensors from the Cense network, chosen to monitor a wide range of scene types, is selected. Over a period of three months, the sensor is given third-octave profiles corresponding to sources of interest. A cosine similarity measure continuously tests if measured third-octave spectra match the given profiles, and any correspondence triggers the recording of a short waveform audio clip. Spectral profiles correspond to subclasses of sound sources of interest found in Lorient: cars, motorcycles, trucks, conversations, crowd noises, laughs, small birds, and seagulls. Particular care is taken to ensure that no intelligible speech can be heard in recorded extracts, and access to the waveform audio is restricted to a few researchers. Each audio clip is listened to and deleted if it may contain intelligible speech, if it does not contain a source of interest or is not monophonic. Otherwise, the extract is trimmed to only the sound event, and in some voice and bird extracts background noise is reduced with state-of-the-art noise reudction techniques available within the Adobe Audition CC software. Obtaining clean extracts representing background activity for sources of interest with the same recording method is difficult, in particular for voice and bird sources. Instead, neutral background noise extracts are taken in all simulated sound scenes. Some extracts are directly obtained from recordings in Lorient, and others are obtained by removing sound events from recordings with Adobe Audition CC software. Thus, in simulated scenes sources of interest are only active during sound events, contrary to *TVBUniversal* where both background and event activity is possible. This yields both positive and negative effects to the training process of deep learning architectures: the network learns to identify sound sources from background noise, but the overall number and complexity of polyphonies in generated scenes is reduced.

The isolated samples database for *TVBCense* is split into a development and evaluation independent subsets containing two-thirds and one-third of available samples, similarly to *TVBUniversal*. Table 4.1 summarizes the distribution of collected extracts and the corresponding total duration for the two subsets. Note that the total duration of background noise extracts is not representative, as in a simulated scene a single extract is taken and

Table 4.1: Contents of the isolated samples database from which simulated scenes in the *TVBCense* dataset are generated.

Subset	Source type	Source class	Extracts	Duration (m)
Development	Background	Neutral noise	16	3:41
	Event	Traffic	128	31:20
		Voice	10	2:22
		Birds	28	3:19
Evaluation	Background	Neutral noise	7	0:54
	Event	Traffic	63	17:16
		Voice	10	2:10
		Birds	17	2:33

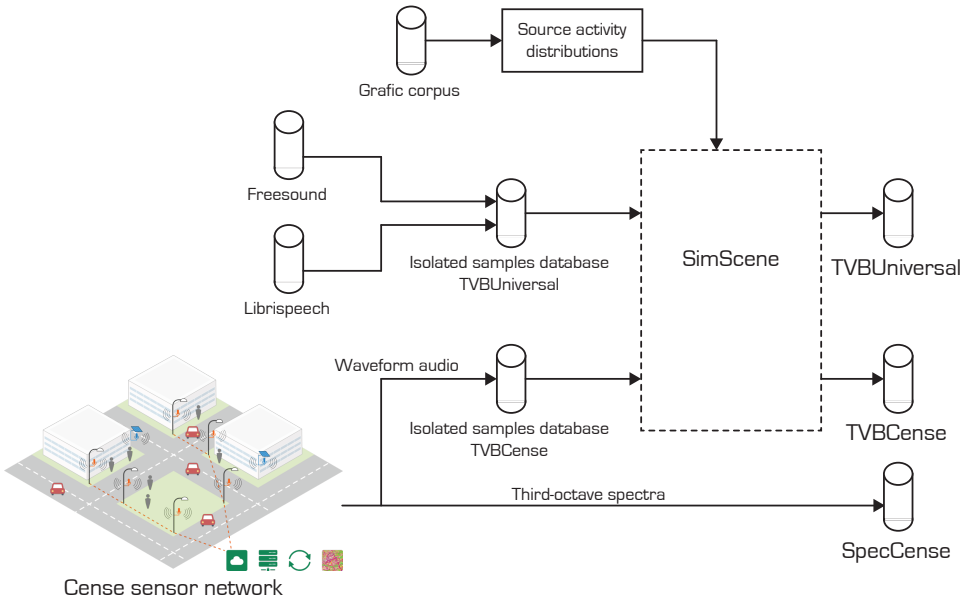


Figure 4.10: Comparison of the construction process for the *TVBUniversal*, *TVBCense* and *SpecCense* datasets in this study.

looped so that it is active for the duration of the scene.

A development set and an evaluation set are simulated from the constructed isolated samples database, together with source activity distributions obtained in Section 2.2.2. The total duration of subsets is identical to that of the *TVBUniversal* dataset. The development set is composed of 400 scenes of 45 s each (total 5 h) and the evaluation set contains 200 scenes of

45 s each (total 2.5 h). Both sets are balanced in terms of ambiances (resp. *quiet street*, *noisy street*, *very noisy street*, *park*, *square*). The *simScene* generation procedure is similar to the construction of the *TVBUniversal* dataset, except that the background source is always characterized by a single extract of neutral background noise. Because background noise extracts contain less signal information and lower energy compared to source specific background extracts in *TVBUniversal*, applying the same event-to-background ratios for sound events results in very salient events in simulated scenes. In particular, event extracts containing some background noise themselves lead to an unrealistic increase to the background noise level for their duration. To palliate this issue and improve the realism of simulated scenes, the mean event-to-background ratios for all event sources are reduced by 6 dB compared to *TVBUniversal*. To compensate the lack of source specific backgrounds and increase the number of polyphonies in generated scenes, the mean event inter-onsets are also empirically divided by a factor of 2, *i.e.* on average twice as many sound events are present in simulated scenes from *TVBCense* than in simulated scenes from *TVBUniversal*. All other distributions input to simScene and described in Appendix A remain unchanged.

All waveform extracts in the isolated samples database are deleted after simulated scenes are generated. Furthermore, after the source presence annotation is obtained waveform simulated sound scenes are also deleted, as they are not useful to train deep learning models. Figure 4.10 summarizes the construction of the *TVBCense* dataset compared to others throughout this study.

4.3.2 Source presence prediction architecture

The best performing model in Chapter 3 on the source presence prediction task is composed of a time-independent convolutional encoder architecture extracting useful information from the input third-octave representation, as well as a recurrent decision process to predict source presence from this information. Thus, a similar model architecture is developed in this chapter. The architecture is shown in Figure 4.11. To enable transfer learning from previous models trained on pretext tasks, the convolutional part of the model in Section 3.3.2 is replaced with the encoder architecture of Section 4.2.2. Sequences of 1 s third-octave texture frames with 875 ms overlap are input to the model, and individual texture frames are processed independently by the encoder architecture. Obtained latent representations are fed to a single-layer gated recurrent unit, that updates an internal recurrent state characterized by a vector of dimension 128 (see Section 3.3.2 for details). A

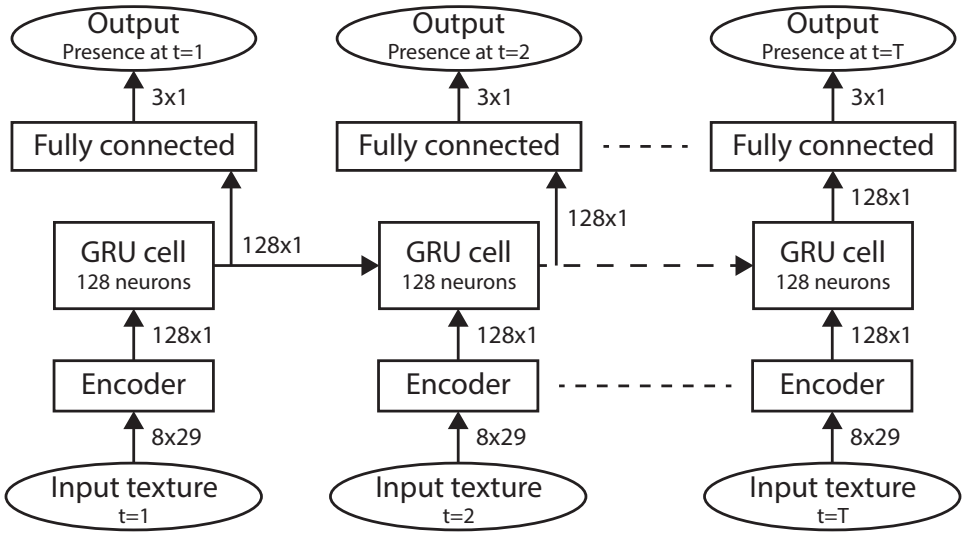


Figure 4.11: Decision architecture for the downstream task of source presence prediciton.

fully connected layer predicts the presence or absence of traffic, voice and bird sources at each timestep of the model, *i.e.* for each input 1 s texture frame, from information in this recurrent state. Model parameters are optimized by minimizing a cross-entropy loss function using the Adam gradient descent algorithm with default parameters. During training, the batch size is set to 64 and the learning rate parameter is set to 10^{-4}.

Several training strategies are possible for transfer learning. First, the entire model can be trained from scratch, that is without first optimizing encoder parameters on a pretext task. This is the reference setting, and should yield comparable performance to models in Chapter 3. Second, the encoder parameters can be initialized with optimal values learned on a pretext task. The encoder can then be frozen during the training procedure on the downstream task. In this case, only parameters of the recurrent decision process are modified. The latent representation output by the encoder can then be regarded as a set of handcrafted features independent of the downstream task model. Because the decision part of the model has a low number of parameters compared to the encoder, the parameters are less likely to overfit on simulated data during training. This also allows us to retain information learned on sensor data in the pretext task, which guarantees good generalization to Lorient environments. However, the low number of parameters and nonlinearities also limit the capacity of the network, which can result in

low performance at convergence if the latent representation is not adapted to the task. Thus, training only the decision architecture has clear advantages in the proposed approach, but requires very close pretext and downstream tasks to yield acceptable performance. Alternatively, the encoder parameters are fine-tuned. In this case optimal parameters learned on pretext tasks replace pseudo-random values as a better initialization. Thus, with three pretext tasks the downstream task architectures can be trained with seven different settings on each simulated dataset.

4.3.3 Evaluation of model performance on target sound environments

Evaluating trained deep learning models on simulated scenes yields limited information regarding their localization capabilities to Lorient environments. Evaluation subsets in simulated datasets are constructed by the same *simScene* process as for training subsets. Thus, they follow the same data distribution and differences compared to sensor data, in terms of missing sources, polyphonies, or ambiances. To better conclude on the generalization of presence prediction models to sensor data in Lorient, this evaluation paradigm is therefore insufficient, and evaluation is done on on-site recordings instead.

To do so, recordings of Lorient sound environments were collected. The recording session took place in one day, and consisted in phases of stationary recording and walking to different locations. Recordings were made with a Zoom h4n microphone as a mobile sensor was not available at this time. As such, there is a discrepancy in terms of frequency response between the microphone in this recording session and acoustic sensors implemented in the Cense network. We assume that the impact of this discrepancy is weak, but should be addressed in further research nonetheless. Every few minutes during the recording session, the A-weighted sound level (in dBA) was measured by a class 1 sound level meter. This allows to scale the level of extracts being processed by the learning models to match sensor measurements in the *SpecCense* dataset. In total, about 2 h of recorded audio in the waveform domain are obtained. 30 non-overlapping scenes of 45 s each are extracted from these recordings (total of 22.5 min). The extracts are only taken during parts of the recording where the person holding the microphone is not walking to avoid sounds of footsteps. Because the recording gain of Zoom microphones varied over the recording session, the scenes are then scaled so that their level difference in dBA matches that recorded by the sound level meter. These sound scenes constitute another dataset referred to as *EvalLorient* in this chapter.

The 30 recorded sound scenes are annotated by a panel of researchers with expertise on acoustics or auditory perception. Two quantities are annotated: the perceived time of presence over 45 s sound scenes and a finer annotation of source activity. Predicting the perceived time of presence is the final objective of trained deep learning models within the framework proposed in Section 1.4. However, the low number of scenes reduces the robustness of this quantity as an evaluation metric. Evaluating trained models on source presence predictions for 1 s third-octave texture frames is thus beneficial, as it allows statistically more robust performance metrics and more thorough investigation of each model's prediction errors. The annotation procedure repeated for each scene is as follows:

1. The sound scene is listened to in its entirety, once and without pausing. The participant annotates a perceived time of presence in the $[0-1]$ range for traffic, voice and bird sources. Background and event sources are not addressed separately.

2. The sound scene is listened to a second time with repeating, pausing and spectrogram viewing freely allowed. The participant annotates finely the scene in terms of activity onsets and offsets for each of the three sources. Annotations are made with 125 ms or more precision, which corresponds to the hop size of frames processed by deep learning models. Furthermore, sound events separated by less than 1 s are merged as deep learning models process texture frames of that duration.

No distinction is made between subclasses of sounds (e.g. small birds and seagulls) to reduce the time needed to complete the annotation process. As participants listen to audio excerpts with their personal computers and headphones, the playback sound levels are not calibrated. Thus, participants are not asked to annotate high-level perceptual attributes (e.g. pleasantness), as they often correlate with the perceived loudness. However, participants are instructed to avoid changing the software audio level during the procedure to conserve correct relative sound levels between sound scenes. Perceived time of presence were annotated by a total of 6 participants, however only 5 among them also completed the fine annotation of sound source activity due to the time consumption involved.

To obtain a single annotation of the perceived time of presence for each source and scene, the assessed perceived time of presence is averaged over participants (n=6). The performance of trained models is then measured by the root mean squared error between the resulting average values and those

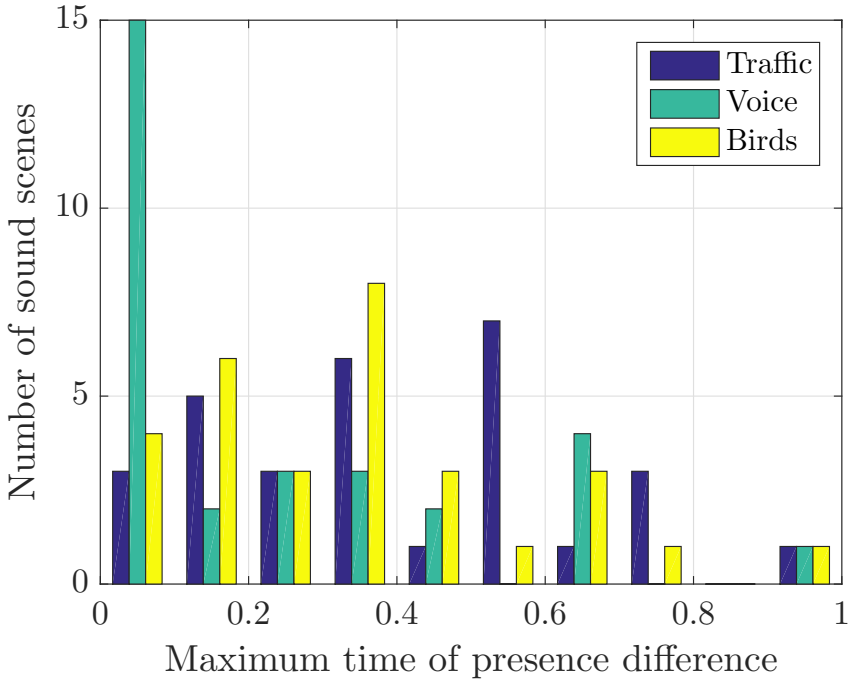


Figure 4.12: Histograms of the maximum difference in time of presence annotations within the same scene among participants.

predicted by the models. This method is valid if the distribution of annotated times of presence has only one mode. However, on some sound scenes high variances in annotated time of presence are observed. In particular, several extracts with background noise resembling traffic activity are perceived as traffic by some participants, yielding time of presence annotations close to 1. Other participants do not perceive these background sources as traffic, resulting in low perceived time of presence annotations in the scenes. This creates a two-mode annotation distribution, thus taking the average as a reference may be ineffective in this case: the average time of presence is close to 0.5 whereas annotations are bimodal with means around 0 or 1. Histograms of the maximum annotation differences for each source are shown in Figure 4.12. Removing the corresponding scenes from the study is difficult as the number of evaluation points is already limited (n=30).

The annotated source activity onset and offset timestamps are quantized with 125 ms precision, *i.e.* they are rounded to the nearest multiple of

125 ms. From these values, a vector of binary source presence for each third-octave 125 ms frame of the sound scene is available for each participant. A unique annotation of source presence is obtained by majority voting, where the source is considered present on a 125 ms frame if the corresponding timestamp is inside an activity onset-offset pair for a majority of participants, and absent otherwise. The presence prediction accuracy of deep learning models can then be simply evaluated by comparing the resulting binary labels to model outputs. However, this metric does not account for annotation variance across participants: annotations are unanimous on 58.8% of labels only, whereas on others they are divided. This phenomenon is particularly observed around the beginning and end of sound events, where it is unclear whether the source should be considered present or absent. To address this issue, a measure of label confidence is proposed that accounts for the unanimity level of labels across participants:

$$Accuracy = \frac{1}{\sum_i w_i} \sum_i w_i \mathbb{1}_{\hat{y}_i = y_i} \tag{4.4}$$

$$w_i = 2 \left| \left(\frac{1}{P} \sum_{p=1}^{P} y_{i,p} \right) - 0.5 \right| \tag{4.5}$$

where w_i is a label confidence value, P is the number of participants in the annotation process, y_i and $\hat{y}_i$ are the annotated and predicted labels for the third-octave frame i, and $\mathbb{1}$ is the indicator function. The confidence is the absolute difference between the mean of annotated binary activity labels and 0.5. Thus, temporal frames for which the source is unanimously labeled present or absent result in an associated confidence score of 1 and contribute fully to the accuracy metric. Conversely, temporal frames for which half of participants considered the source present and the others considered the source absent yield a confidence score close to 0 and are omitted in the computation of accuracy. On the 30 scenes in the *EvalLorient* dataset, the average confidence score of labels is 0.73.

4.4 Experiments

4.4.1 Evaluation of the transfer learning approach

A first set of experiments is conducted to assess the interest of learning audio representations on pretext tasks in order to localize source presence prediction tools. The proposed deep learning architecture for the downstream

Table 4.2: Source presence prediction accuracy (%) of models with encoder parameters pre-trained on pretext tasks compared to learning from scratch. The three metrics are respectively the accuracy on the *TVBCense* evaluation subset, the standard accuracy on the *EvalLorient* corpus, and the accuracy modified with label confidence on the *EvalLorient* corpus.

Encoder training	TVBCense	EvalLorient	EvalLorient (Conf.)
Scratch	82.40	76.90	83.30
SensorID, frozen	79.17	67.22	70.40
SensorID, finetuned	79.27	74.90	82.73
Time, frozen	80.81	65.75	68.42
Time, finetuned	81.66	71.83	79.61
Audio2Vec, frozen	78.67	66.12	65.88
Audio2Vec, finetuned	84.17	76.89	83.31

task is trained on the *TVBCense* dataset with the 7 settings discussed in Section 4.3.2. These settings include training both the encoder and the decision from scratch, and using either frozen or fine-tuned encoder parameters pre-trained for each of the *SensorID*, *Time* and *Audio2Vec* pretext tasks. Accuracy metrics for source presence prediction on individual third-octave texture frames are first discussed. Table 4.2 compares the performances of trained models on the *TVBCense* evaluation subset, as well as on recorded scenes in the *EvalLorient* dataset. Both the standard accuracy and the modified accuracy that includes a measure of label confidence (see Section 4.3.3) are compared.

On the *TVBCense* evaluation dataset, the highest accuracy (84.17%) is obtained by pre-training the encoder architecture on the *Audio2Vec* task. This setting slightly outperforms training from scratch (82.40%) as well as other pre-training settings (around 80%).

The accuracy of the best model on the *TVBCense* evaluation set is also lower than that of models trained in Chapter 3 on the *TVBUniversal* dataset (93.67% with a recurrent architecture), by about 10%. This is mostly due to a high false positive rate on traffic activity detection, which is at 46.7% in the model trained from scratch and similar in all settings involving pre-trained encoders with finetuning. Sound scenes in *TVBCense* are simulated with a neutral background noise source instead of source-specific background extracts in *TVBUniversal*. This background noise is often interpreted as traffic by the models. To help reduce this issue, the training procedure is repeated with an added class corresponding to neutral background noise. This forces

models to identify background noise sources from events. Presence labels are computed in the same way as for the three sources of interest using the indicator developed in Section 2.4. However, this method yields slightly worse overall performance in all model training settings.

On both accuracy metrics computed on Lorient recordings in the *EvalLorient* dataset, pre-training encoder parameters on the *Audio2Vec* and fine-tuning them in the downstream task learning process leads to the same performance as the model trained from scratch (83.31% to 83.30%). As expected, the accuracy is higher for all models when label confidence is introduced. Discriminative pretext tasks yield worse performance than training from scratch with both frozen and fine-tuned encoder parameters. Overall, pre-training encoder parameters does not increase model accuracy in the current setting. This result may be due to two factors: the domain of *TVBCense* matches well that of sensor data, or the information content of learned latent representations to be relevant for solving the target task, *i.e.* pretext tasks are not sufficiently related to source presence prediction.

Freezing encoder parameters during the downstream task training process also results in worse performance compared to fine-tuning them in all cases. As discussed in Section 4.1.4, fine-tuning encoder weights in the proposed approach compromises between improving the capacity of the deep learning architecture and losing information learned on sensor data during the training of pretext task, that is not contained in simulated scenes. In other terms, there is a point where freezing encoder parameters is beneficial if the downstream task dataset domain is too different from that of sensor data, and if the pretext task is closely related to the downstream task. Thus, this result may indicate that the content and diversity in the *TVBCense* simulated dataset is sufficiently close to that of sensor data, so that the loss of information learned in the pretext tasks is outweighted by the benefits of increased model capacity.

Next, the final task objective is investigated, by studying the error in perceived time of presence estimations compared to human annotations. Estimations of the time of presence are obtained by averaging predictions of deep learning models over time. Trained models are evaluated on the root mean squared error (RMSE) metric on a $[0-1]$ scale, and results are shown in Table 4.3. Similarly to local presence prediction accuracy, the model with an encoder pre-trained on the *Audio2Vec* task performs as well as the model trained from scratch. Other settings with finetuning of pre-trained encoder parameters result in higher time of presence errors across all sources. Interestingly, models with frozen encoder parameters always yield better estimates of the time of presence of traffic compared to their finetuned counterpart.

Table 4.3: Time of presence prediction root mean squared errors on a [0-1] scale yielded by predictions of proposed models on Lorient recordings (n=30).

Encoder training	All sources	Traffic	Voices	Birds
Scratch	**0.26**	0.36	0.20	0.19
SensorID, frozen	0.39	0.37	0.41	0.39
SensorID, finetuned	0.31	0.44	0.23	0.20
Time, frozen	0.39	0.35	0.50	0.28
Time, finetuned	0.33	0.46	0.28	0.20
Audio2Vec, frozen	0.35	0.26	0.46	0.28
Audio2Vec, finetuned	**0.27**	0.36	0.23	0.21

However this performance is balanced by poorer estimations for voice and bird sources. Overall high errors are obtained, especially compared to results in Chapter 3 obtained from annotations in a listening test. This is in part explained by high annotation variance among participants as discussed in Section 4.3.3, where bimodal annotation distributions limit the validity of perceived time of presence averages in some sound scenes.

Pre-training encoder architectures on pretext tasks does not improve source presence detection or time of presence estimations on the *EvalLorient* dataset compared to training from scratch. We hypothesize that this result is due to a combination of two conditions. First, the domain of *TVBCense* is sufficiently close to that of sensor data. Thus, the large dataset of sensor data provides limited knowledge that is not also contained in *TVBCense*. Secondly, with a total of 5 h of audio the downstream task dataset is sufficiently large so that trained models generalize well. In contrast, domain adaptation problems are generally investigated for very low amounts of downstream task data available. The following experiments thus aim at testing the two parts of this hypothesis in order to determine the conditions under which pre-training is beneficial.

4.4.2 Content of simulated training datasets

The *TVBCense* dataset is simulated from a localized isolated samples database with recordings made in target environments. To assess the impact of this localization and resulting higher domain proximity with sensor data, *TVBCense* is compared with the *TVBUniversal* dataset designed in Chapter 3. Although both datasets have the same total duration of 5 h, the isolated samples database in *TVBUniversal* contains samples obtained from online repos-

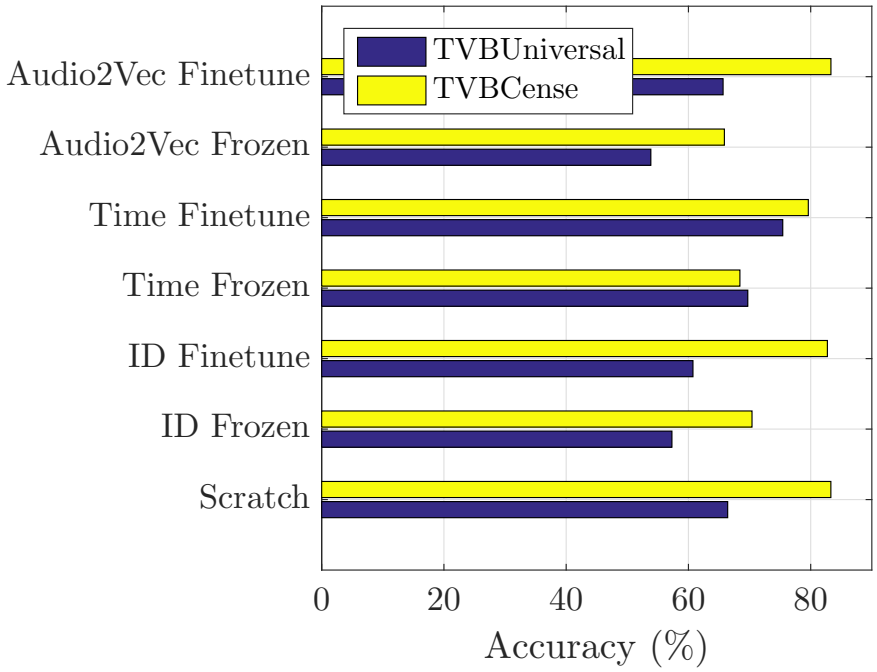


Figure 4.13: Comparison of presence prediction accuracy in *EvalLorient* recorded scences between models trained on the *TVBUniversal* and *TVBCense* simulated datasets.

itories. Deep learning models are trained on the *TVBUniversal* development subset and evaluated on the source presence prediction modified accuracy for scenes recorded in Lorient (*EvalLorient*). Figure 4.13 compares the accuracy of models trained on the two datasets of simulated scenes. When training the downstream task model from scratch, *TVBUniversal* achieves poorer accuracy (66.42%) compared to *TVBCense* (83.30%). Large differences are also found in settings where the encoder architecture is pre-trained, with the exception of the *Time* pretext task. Both training settings with this pretext task outperform training from scratch, and the model with finetuned encoder parameters achieves 75.43% accuracy. In contrast, the benefits of informative latent representations obtained with the *Audio2Vec* and *SensorID* tasks are not well exploited by the downstream task training process on the *TVBUniversal* dataset.

These results confirm the effectiveness of matching content in the down-

stream task training dataset to target sound environments, in order to improve the localization of prediction tools. If obtaining local recordings of isolated source occurences is not possible, pre-training encoder parameters on the *Time* pretext task improves accuracy, by about 9% in this experiment.

4.4.3 Quantity of downstream task training data

In a typical transfer learning setting, representation learning on pretext tasks is particularly useful when very low amounts of data annotated for the downstream task are available. Experiments are conducted to assess whether a similar behavior is observed in the context of localizing architectures trained on simulated data. In datasets of simulated scenes, intrinsic data quantity mainly depends on two factors. First, the diversity of spectral patterns associated to sources of interest, which contributes to the generalization of trained deep learning models, depends on the size and sample diversity of source-specific extracts in the isolated samples database. Second, the diversity of scenarios and polyphonies is linked to the number of sound scenes simulated in the training dataset. The effects of these two factors are partly correlated. Increasing the number of sound scenes improves the diversity of polyphonies and scenarios in the dataset, although the maximum diversity attainable is dictated by that of the isolated samples database. Similarly, the effect of increasing the size of the isolated samples database is limited if the simulated dataset is too small to contain polyphonies with all source extracts.

The impact of varying the number of sound scenes is first studied, *i.e.* the diversity level of the isolated samples database taken to generate the simulated scenes remains unchanged in all experiments. To do so, experiments are conducted where the number of sound scenes in the training subset of the downstream task is reduced. Specifically, models trained on the full *TVB-Cense* development set (5 h) are compared to models trained on datasets reduced to 30 min and 1 h of data respectively. The reduced datasets are obtained by selecting only 40 and 80 of the 400 simulated scenes in the full development set. These subsets are balanced in terms of ambiances (*quiet street*, *noisy street*, *very noisy street*, *park* and *square*). Performances are evaluated on the modified accuracy metric for the *EvalLorient* recorded scenes. Figure 4.14 shows the evolution of the presence prediction accuracy as a function of increasing dataset size.

With a dataset duration of 1 h, both models pre-trained on the *Audio2Vec* and *Time* pretext tasks slightly outperform the model trained from scratch. The difference is amplified for a dataset of 30 min, where pre-training on

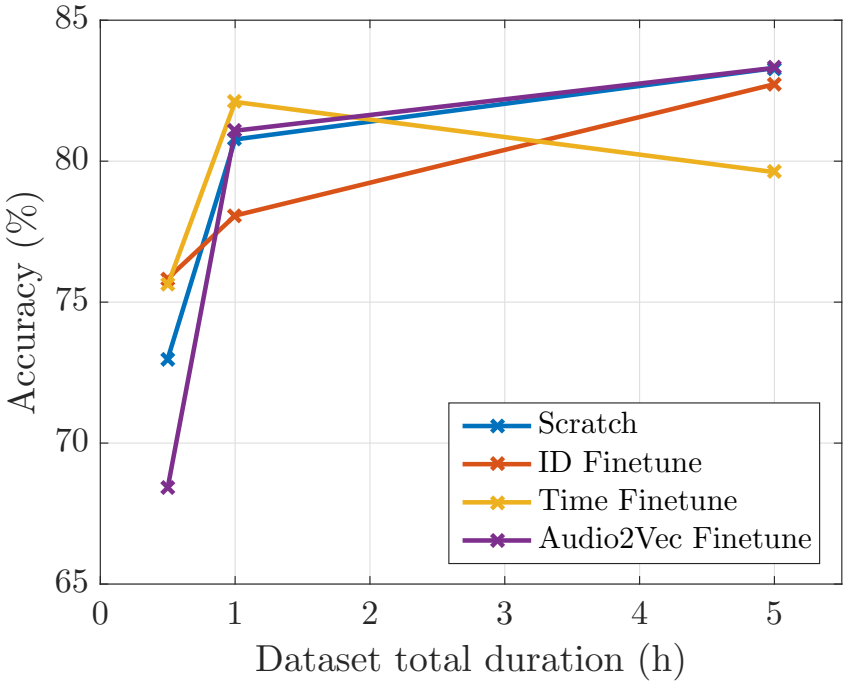


Figure 4.14: Performance of models trained with a limited number of simulated scenes evaluated on source presence accuracy in Lorient recordings.

Time and *SensorID* yields 75.61% and 75.84% accuracy respectively, whereas the accuracy of the model trained from scratch drops to 72.99%. The model trained on the full *TVBCense* set with pre-training on the *Time* task achieves lower accuracy compared to that trained on only 1 h of simulated data. This points to the possible overfitting of model parameters to properties specific to simulated data. This type of overfitting may not be detected during the training process due to validation examples being also simulated.

For total dataset durations of over 5 h, accuracy metrics should converge to a maximum value associated to each encoder training setting. This value is close for settings where the entire model is trained from scratch and where the encoder is pre-trained on the *Audio2Vec* or *SensorID* pretext tasks, at around 83%. This difference illustrates the distance between distributions of sensor data and simulated environments. Thus, a small difference further indicates that the diversity of scenarios and spectral patterns associated to sources of interest in simulated data, although limited, represents well the

corresponding diversity found in acoustic environments in Lorient.

In practice, simulation tools paired with automatic annotation procedures allow the generation of arbitrarily large number of labeled sound scenes with different scenarios. The main factor limiting information content in simulated datasets is thus the richness of the associated isolated samples database. The second experiment thus evaluates models trained on datasets with reduced isolated samples databases in order to determine wether pre-training is beneficial when the number of available extracts is low. Specifically, two settings are addressed where the number of samples corresponding to each class in Table 4.1 is divided by 2 and 4 respectively. For example, the number of traffic event extracts is reduced to 64 and 32 in the two settings respectively. Instead of considering equal number of samples for each class, this approach is motivated by the higher difficulty of obtaining clean samples for some sources, in particular human voice. The number of background noise extracts is also reduced as they are similarly obtained from local recordings.

Because the number of generated sound scenes is not limited, all training datasets contain 5 h of audio. Figure 4.15 shows model accuracy on *EvalLorient* as a function of the isolated samples database reduction ratio. Interestingly, models pre-trained on the *Audio2Vec* task respond best to lower diversity in simulated datasets, with an accuracy of 80.36% and 75.52% with isolated samples databases divided by 2 and 4 respectively. In contrast, the accuracy of the model trained from scratch increases almost linearly as a function of the isolated samples dataset size (resp. 73.11%, 76.46% and 83.30%). All settings including encoder pre-training outperform the model trained from scratch on the simulated dataset with an isolated samples database reduced by a factor of 4. Further reducing the isolated samples database is difficult in the current experiment as the smallest only contains 3 voice extracts.

The two experiments demonstrate the interest of pre-training encoder parameters when fewer training examples are available, or when they contain less diverse spectral patterns. In the first case, discriminative pretext tasks are preferable, although it is not likely in practice due to the ability of simulation tools to generate arbitrarily large datasets. In the second case the *Audio2Vec* generative pretext task performs better overall. Retaining a large number of training examples in the downstream task is necessary for pre-training with the *Audio2Vec* pretext task to be beneficial, compared to the two discriminative tasks. This behavior may be linked to the information content of corresponding latent audio representations which should be higher in the *Audio2Vec* pretext task, thus making downstream training prone to

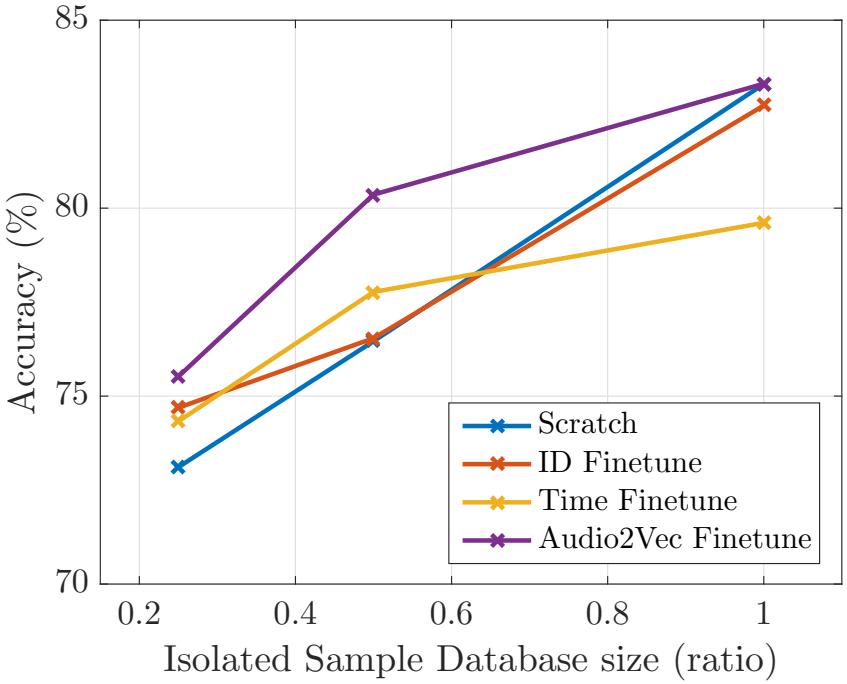


Figure 4.15: Performance of models trained with a limited isolated samples database evaluated on source presence accuracy in Lorient recordings.

overfitting with lower amounts of data.

In the investigated setting, freezing learned encoder parameters always leads to worse performance. This is possibly due to the high number of labeled examples in simulated datasets, which mitigates overfitting issues typically encountered in domain adaptation problems where only few training examples are available.

Chapter conclusion

The adaptability of deep learning models trained on location-independent simulated datasets (Chapter 3) to in situ sound environments of Lorient is limited. Experiments conducted in this chapter demonstrate that the localization properties of simulated corpora are an important factor of the performance of predictive models, and including on-site recordings of sources in their construction increases prediction accuracy on evaluation recordings. In some cases, transfer learning techniques enable the domain specialization of

trained models, by learning well-generalized latent audio representations on large amounts of sensor data. Specifically, discriminative pretext tasks such as *Time* can mitigate the domain mismatch between location-independant datasets and sensor data, and improve model adaptability when few labeled examples are available. Conversely, the generative *Audio2Vec* pretext task outperforms other settings when spectral patterns of sources in localized simulated scenes are less diverse. Work still remains to enrich simulated corpora with additional location-specific scenarios and sound sources. Regarding the pretext tasks, future work should consider enlarged training datasets both in terms of spatial diversity and time span. Nonetheless, the developed approach can be implemented as presented within the Cense sensor network.

Chapter 5

Acoustic scene synthesis from sensor features

The ability to listen to audio extracts corresponding to sensor measurements is beneficial to citizen information and urban design. In the current application, sensors measure log-frequency spectral representations with information content limited by privacy constraints. Recovering sound scene examples from such representations is an open problem in the deep learning community. Here, two deep learning approaches are developed to approximate the inversion of the non-invertible third-octave transform to short-term Fourier transform magnitudes.
The first approach is a deterministic architecture which builds upon an initial estimation obtained by application of a third-octave transform pseudoinverse, by inferring a corrective term. This method improves reconstruction quality on objective metrics, but does not recover fine structures. The second model is designed to better account for the polynomial decay of environmental sound spectra with respect to frequency, which causes regression loss functions to ignore errors in high-frequency content prediction. To do so, separate architectures produce content in different frequency bands in a limited weight sharing paradigm. Furthermore, the model is stochastic and trained in an adversarial setting to enable the production of relevant spectral variations in generated examples.

5.1 Introduction

In accordance to privacy regulations, sensors implemented as part of the Cense project record limited information about sound environments. Specifically, the information content of measurements should not allow the reconstruction of intelligible speech [18]. To do so, sensors measure spectral representations through non-invertible signal processing operations, that are yet sufficient to infer acoustic indicators in traditional monitoring applications. This chapter investigates synthesizing waveform audio examples from these privacy-aware representations.

The main motivation of this contribution is to ultimately provide citizens, city administrators, and urban planners with easily interpretable information about the sound environments associated to continuous spectral measurements, along with perceptual attributes discussed in previous chapters. Although perfect reconstruction of waveform audio is not necessary to this aim, synthesized sound scenes should contain recognizable sources and convey relevant perceptual properties about the represented type of environment.

Sound synthesis, and particularly in strongly conditioned contexts where high-dimensional information is available, is a difficult task and an open problem in the deep learning community. Although methods are now being actively developed for synthesizing speech and music signal [109, 110], applications rarely address environmental sounds. Compared to music and speech sounds, environmental sounds have specific characteristics, often with a high level of polyphony as well as very diverse spectral patterns produced by a large range of sound sources. As such, the associated synthesis task is highly complex and requires important modeling capabilities in deep learning approaches.

In the applicative context of the Cense sensor network, measurements are third-octave energies in the range of audible frequencies, with a step size of 125 ms between non-overlapping audio frames. This design introduces two types of information loss to account for during resynthesis, one along the time axis and the other along the frequency axis. First, there is no information redundancy between adjacent analysis frames in sensor measurements. Such redundancy is necessary to the inversion of spectrograms to waveform audio with the overlap-add method, as well as to guarantee the convergence of phase recovery iterative algorithms. Secondly, applying a third-octave transform summarizes spectral magnitudes on logarithmically spaced frequency bands. In this chapter, we choose to focus only on the recovery of the information lost due to third-octave analysis, and leave the interesting issue

of temporal interpolation to future work. To do so, we consider a simplified setting where temporal analysis frames overlap.

5.2 Related work

The task of inverting informative spectral features to realistic waveform audio has been extensively investigated in the literature. In particular, the current problem of recovering audio from third-octave spectrograms is similar to the vocoding of Mel spectrograms or derived Mel-Frequency Cepstral Coefficients (MFCC) in the speech processing community. Traditional signal processing approaches consider approximate inversion of the application of Mel filterbanks, which can be formulated as matrix multiplication in the spectral domain. To do so, a pseudoinverse of the forward transformation matrix can be computed [111]. However, this inverse approximation only depends on the filterbank and does depend on a priori information on the properties of considered audio. In subsequent works, general properties of speech signals are taken into account by introducing constraints on the Mel pseudoinverse estimation. For example, in [112] the optimization process of a pseudoinverse of the Mel filterbank matrix is subject to non-negativity constraints in order to guarantee the non-negativity of recovered fine-band spectra. The approach developed in [113] is based on minimizing a L_1 objective function to better match the sparsity of clean speech spectra, instead of the typical least squares setting for which the optimal solution is the pseudoinverse.

Recent work in spectral feature inversion and general sound synthesis is mainly based on deep learning approaches. In spectral sound synthesis approaches, time–frequency representations are considered as images, and architecture designs are typically borrowed from the image synthesis community. Iterative algorithms then recover the remaining phase component necessary to invert spectrograms to waveform audio. The most successful approaches in recent literature on image synthesis are based on generative adversarial networks (see Section 5.5.1 for details). An adversarial architecture is proposed in [114] to solve the task of unconditioned synthesis of short-term Fourier transform (STFT) magnitude spectrograms. A similar approach is applied in [115] to the inversion of speech Mel spectrograms, with the *AdVoc* auto-encoder generator architecture conditioned on the pseudoinverse estimation of STFT magnitudes. Section 5.3.3 further presents this architecture which constitutes a baseline to in comparison to the developed approach.

By modeling waveform audio signals directly, sample-based approaches

currently achieve state-of-the-art performance both in speech and music generation tasks. These methods include likelihood-based autoregressive models such as WaveNet [54, 116] and SampleRNN [117], which model the conditional distribution of each audio sample given past samples by modeling temporal structures at different rates. In particular, a WaveNet architecture vocodes waveform speech from generated Mel spectrograms for text-to-speech in [109]. However, the capability of autoregressive architectures to model long-term structures in audio signals while synthesizing audio with high sample rates generally depends on the number of parameters and overall computational complexity of the model. Other developed sample-based approaches perform better in terms of long-term structure learning for weakly conditioned or unconditioned synthesis, including flow-based models [118] and variational auto-encoders (VAE). The current state of the art in music synthesis is achieved in [119], where transformer architectures [120] condition vector-quantized VAE decoders modeling waveform music at different levels of abstraction (time scales). The trained architecture is able to capture long-term structure up to a few seconds, but totals several billion parameters resulting in important data and computation capability needs.

Recently, the authors of [110] have proposed the Differentiable Digital Signal Processing (DDSP) framework, that allows to include either exact or differentiable approximations of signal processing operations in the training process of deep learning architectures. This represents a shift from paradigms considered in the best performing sample-based approaches. Models such as Jukebox [119] rely on very large number of model parameters and conventional deep learning layers regarded as universal function approximators to learn increasingly complex patterns and long-term structure of sounds. In contrast, DDSP allows the development of architectures well motivated by the task objective, with the aim of providing more interpretable models as well as higher parameter efficiency. This approach is thus a promising alternative to large sample-based models in future studies.

There is thus in recent approaches a tradeoff between information content in synthesized audio and model complexity. Spectral approaches require less parameters and data to train, but do not account for phase information. Sample-based approaches circumvent phase recovery by modeling waveform signals directly, but they are dependent on very large architectures and datasets to synthesize audio at sufficiently fine scales while accounting for long-term patterns.

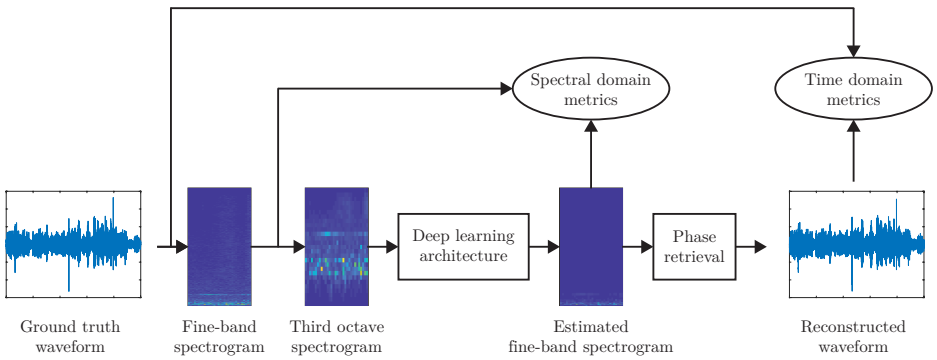


Figure 5.1: Proposed spectral approach to waveform reconstruction from third-octave spectrograms. A deep learning architecture outputs an estimation of the fine-band spectrogram, which an iterative phase recovery algorithm inverts to waveform audio.

5.3 Experimental protocol

5.3.1 Considered approach

Due to the computational complexity of sample-based approaches, this study instead addresses spectral feature inversion. Figure 5.1 summarizes the method, where a deep learning model reconstructs plausible short-term Fourier transform magnitude spectrograms from input third-octave texture frames. The phase component is then recovered by iterative algorithms further described in Section 5.3.4.

5.3.2 Dataset

Training deep learning models for the task of spectral feature inversion requires the availability of reference spectrograms associated to third-octave examples. This information is not collected by sensors in the Cense network. Instead, architectures presented in this chapter are trained on the popular and publicly available TUT Acoustic Scenes 2017 dataset, proposed as part of the DCASE2017 challenge task 1 [56]. This dataset of environmental sound scenes is referred to as the DCASE2017 dataset in the rest of this chapter.

The DCASE2017 dataset is composed of two independent subsets for development and evaluation, with a total of 13 h and 4.5 h of waveform audio respectively. It is constructed from indoor and outdoor recordings collected

in Finland between 2015 and 2017. The recordings of several minutes are split in 10 s segments corresponding to individual sound scenes. Sound scenes extracted from the same recordings are always placed in the same subset to ensure the independence of the evaluation set.

The DCASE2017 dataset is designed for acoustic scenes classification tasks. Thus, each sound scene is annotated with an ambiance among 15 classes (*beach*, *bus*, *cafe/restaurant*, *car*, *city center*, *forest path*, *grocery store*, *home*, *library*, *metro station*, *office*, *park*, *residential area*, *train*, *tram*). Both the development and evaluation subsets are balanced in terms of ambiances. Furthermore, four independent and balanced splits of the development dataset are provided for cross-validation purposes. For each split, parameter optimization is done on 75% of the subset and validation is done on the remaining 25% of examples. All experiment results presented in this chapter are obtained with the first split configuration.

Sound scenes in the DCASE2017 dataset are pre-processed to obtain training and evaluation examples. In order to ensure the validity of comparative studies with the *AdVoc* deep learning baseline considered in Section 5.3.3, audio pre-processing steps are designed so that its architecture needs not be modified. All sound scenes are thus resampled to 22.5 kHz, and fine-band spectrograms are extracted by applying a short-term Fourier transform on frames of 1024 samples (46.4 ms) with 75% overlap. High information redundancy between adjacent frames is in accordance with the proposition in Section 5.1 to only focus on the frequency degradation in third-octave spectra. Third-octave spectra are computed for each 46 ms frame for bands with center frequency in the 20 Hz to 8 kHz range (total of 26 bands). All objective metrics and loss functions in this chapter are computed within this range. Individual examples are constructed as groups of 256 adjacent frames, which corresponds to about 3 s of audio data. These examples are processed independently by the *AdVoc* baseline as well as the proposed deep learning models.

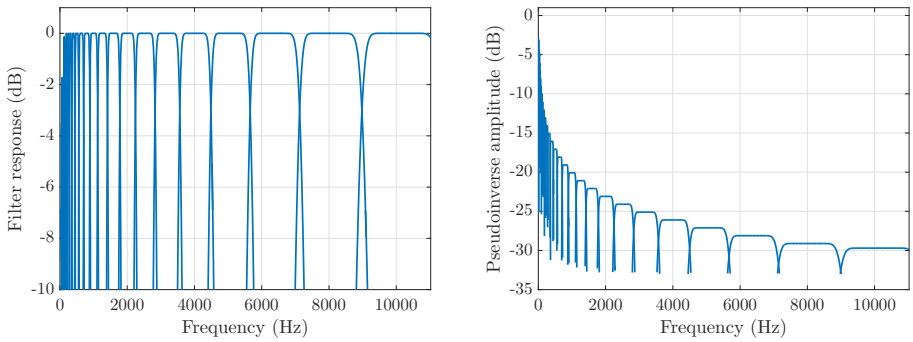
5.3.3 Baselines

Two baseline methods of spectral feature inversion proposed in the literature are considered. The first baseline only relies on the properties of the third-octave transform to estimate short-term Fourier transform magnitudes. The third-octave analysis consists in applying a filterbank to an input waveform audio example, and summarizing the energy of the filtered signal over temporal analysis frames. This transformation can be formulated in the Fourier

domain as a matrix multiplication:

$$X = \Phi X_f, X_f = |\mathcal{F}(x)|^2 \tag{5.1}$$

where X is the third-octave spectrogram with dimensions $B \times T$, Φ is the third-octave filterbank matrix with dimensions $B \times F$, X_f is the fine-band spectrogram of dimensions $F \times T$, x is the input waveform audio and $\mathcal{F}$ denotes the short-term Fourier transform. Since the number of third-octave bands B is lower than the number of fine bands F, the transform matrix Φ has no left inverse. However, a Moore-Penrose pseudoinverse $\Phi^\dagger$ exists, and its application to the third-octave spectrogram X produces an estimate of X_f. Pseudoinverse approximations can be computed using either a least squares solver or the singular value decomposition (SVD) of the forward transformation matrix Φ. Both methods yield almost identical results in the current study. However, pseudoinverse computation algorithms are not subject to non-negativity constraints. A threshold $\hat{X} = max(0, \hat{X})$ further sets negative negative spectral magnitude estimations to 0. This results in a lower estimation error than imposing a non-negativity constraint to the $\Phi^\dagger$ matrix directly.

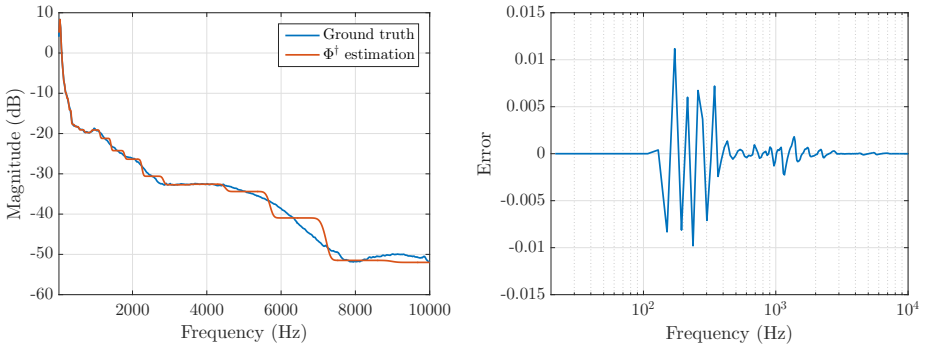


(a) Third-octave filterbank (b) Pseudoinverse

Figure 5.2: Frequency response of third-ocrave filters in the $[20Hz - 12500Hz]$ range (a), and weights of the associated pseudoinverse matrix (b).

The performance of pseudoinverse baseline is intrinsically limited by the fact that it is independent from the content of processed data. Thus, it uses no prior knowledge about the spectral energy distribution in environmental sounds, or about the relation between temporal frames. Third-octave filters have a flat frequency response between cutoff frequencies, as shown in Figure 5.2(a). The pseudoinverse in Figure 5.2(b) therefore assumes that

the magnitude of the original signal is flat in the corresponding frequency range. In other terms, the pseudoinverse approach is optimal in the case of white noise signals. Figure 5.3(a) shows an example of environmental sound average spectrum and its reconstruction by the pseudoinverse transform. Environmental sounds typically display inversely polynomial decrease in power as a function of frequency, close to 1/f [121]. This property results in high variations of the estimation error at filter cutoff frequencies, whereas the error is on average close to 0 around the center frequency of third-octave bands. This behavior is illustrated in Figure 5.3(b). In contrast, applications of this method on speech signals in the literature performs well because of the relative limited bandwidth occupied by speech information, as well as the shape of filters in other filterbanks (e.g. Mel filters).



(a) (b)

Figure 5.3: Average power spectrum of acoustic scenes in the DCASE2017 development dataset and its reconstruction by the pseudoinverse transform (a). The estimation error resembles a sawtooth waveform due to flat estimations within filter bandwidths (b).

As a second baseline, the Adversarial Vocoder (*AdVoc*) deep learning approach is considered [115]. This method is initially proposed for the inversion of Mel spectrograms and applied to speech signals. Input Mel spectrograms are constructed from waveform audio sampled at 22 050 Hz, and consist in texture frames of Mel spectra computed on 46 ms frames with 75% overlap. Mel filters are only in the $[125Hz, 7600Hz]$ range, in which nearly all energy from speech signals is contained. Input examples are groups of 256 frames, about 3 s of audio signal. The pseudoinverse of the Mel transformation is first applied to input spectrograms, yielding an estimation of the associated fine-band spectrogram. This estimation is then processed by an encoder-decoder

generator architecture. The encoder is composed of 8 convolutional layers that downsample the input representation to a vector of dimension 512. The decoder mirrors the encoder with 8 transposed convolutions that upsample the embedding to the initial spectrogram dimensions (see Section 5.5.2 for details). The generator is trained in an adversarial setting: a discriminator model jointly learns to classify real and fake (generated) examples, with the objective function in eq. 5.7. The discriminator provides a gradient to the generator so that the distribution of generated samples converges towards that of real samples. Additional detail about the training of generative adversarial networks is discussed in Section 5.5.1. Furthermore, training on an adversarial loss only forces the generator to produce realistic examples, *i.e.* examples that match the distribution of data in the training set. Because the synthesized fine-band spectrogram should also correspond to the specific input example, an additional L_1 loss term is minimized between the generator's output and the ground truth fine-band spectrogram associated to the input. Parameters of the generator are thus optimized to minimize a combination of two losses, respectively influencing the realism and correspondence with the ground truth of synthesized spectrograms. The phase component necessary to invert generated magnitude spectrograms to waveforms is obtained with an iterative optimization method, the Local Weighted Sums (LWS) algorithm proposed in [122]. This algorithm is briefly described in Section 5.3.4.

No modification of the implementation proposed by the authors is necessary in order to apply the *AdVoc* model to the current problem, as the same audio pre-processing parameters are considered. Yet, input third-octave spectrograms with 26 bands are zero-padded in the frequency dimension to match the size of the expected input representation (Mel spectrograms with 40 filters) in the original implementation. Experiments also retain the optimization method and hyper-parameters proposed by the authors.

5.3.4 Phase recovery

The proposed approach consists in reconstructing fine-band magnitude spectrograms from third-octave measurements. In order to produce audio in the waveform domain, the phase component also lost during analysis must be recovered. To do so, the first method applies the original phase of the audio signal before third-octave analysis. This method is termed the *oracle* phase in experiments, as this information is typically not available in the application. However, it allows the comparison of synthesized sound scenes with the corresponding ground truth on phase-sensitive metrics, in particular metrics

involving sample-wise waveform errors such as the signal-to-reconstruction ratio presented in Section 5.3.5.

In a context where ground truth phase information is not available, phase recovery algorithms can recover estimations the magnitude spectrogram. Waveform samples in the *AdVoc* baseline are obtained with the Local Weighted Sums (LWS) algorithm [122]. This method is based on the same principles as the Griffin-Lim algorithm [123], *i.e.* it iteratively updates an estimate of the phase component to maximize the consistency of the resulting waveform signal y, where consistency is given by the distance between y and ISTFT(STFT(y)) (resp. the inverse and forward short-term Fourier transform). This method achieves similar performances to the Griffin-Lim algorithm, although with faster convergence. To remain comparable with the *AdVoc* baseline, phase estimations for all discussed methods are obtained with 60 iterations of the LWS algorithm.

5.3.5 Objective metrics

The investigated synthesis task is strongly conditioned with input third-octave spectrograms, each associated with ground truth waveform audio examples in the dataset. The content of generated spectra should match that of corresponding ground truth sound scenes. This differs from weakly conditioned synthesis tasks, for example the synthesis of realistic spectrograms representative of an ambiance with no constraint on specific scenarios and active sources, where only the realism and correspondence to a conditioning class label are evaluated. It is thus necessary to assess model performance on two separate qualities: i) the correspondence between generated and ground truth spectrograms, and ii) the presence of sufficient fine-grain detail in synthesized spectra to identify sound sources as well as the type of environment.

The performance of models in terms of correspondence between synthesized and ground truth sound scenes is evaluated on signal reconstruction metrics. The Signal-to-Reconstruction Ratio (SRR) is first considered as a waveform domain metric. The SRR is computed between a synthesized signal $\hat{y}(t)$ and a target signal $y(t)$ as:

$$SRR(\hat{y}, y) = 10 \log_{10} \left(\frac{\sum_t y(t)^2}{\sum_t (\hat{y}(t) - y(t))^2} \right) \tag{5.2}$$

Comparing a signal with itself yields $SRR = +\infty$, and the SRR decreases as the mean squared error $(\hat{y}(t) - y(t))^2$ increases. In the current experiments, the SRR is computed for time-domain signals obtained using both the oracle phase and the Local Weighted Sums (LWS) algorithm. However, because

it involves the sample-wise difference between the synthesised and ground truth signals, the SRR is extremely phase-sensitive, and is not expected to yield reliable performance evaluation with approximate phase recovery.

The Log-Spectral Distance (LSD) is considered as a reconstruction metric in the spectral domain. The LSD is defined as the mean squared error (L_2 distance) between log-power synthesized $\hat{Y}(t,f)$ and ground truth $Y(t,f)$ spectrograms:

$$LSD(\hat{Y},Y) = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{F}\sum_{f=1}^{F}\left(\log_{10}\frac{Y(t,f)^2}{\hat{Y}(t,f)^2}\right)^2} \tag{5.3}$$

where T and F are the time and frequency dimensions of spectrograms respectively. The LSD is positive, and reaches 0 for two equal power spectra.

Some metrics have further been proposed that correlate well with human assessments of reconstruction quality. Here, the Perceptual Similarity Metric (PSMt) is considered, which is based on a model of peripheral auditory processes and proposed as part of the PEMO-Q method [124]. The PSMt metric is computed between generated and ground truth waveform signals using the implementation available in the PEASS software library [125, 126].

The second evaluated aspect is the ability of synthesized audio examples to realistically convey perceptual attributes specific to the type of environment. To do so, the Inception Score (IS) is considered. The IS is proposed in [127] as an evaluation metric for generative adversarial networks. The metric relies on the availability of class labels associated with dataset examples, as well as a classifier model C trained to accurately predict these classes from real examples. It is defined as:

$$IS = exp\left(\frac{1}{N}\sum_{n=1}^{N} D_{KL}(p(y_C|x_n)||p(y_C))\right) \tag{5.4}$$

where x_n is a generated example, N is the total number of generated examples, D_{KL} is the Kullback-Leibler divergence, $p(y_C|x_n)$ is the probability distribution of classes predicted by the classifier C, and $p(y_C)$ is the marginal class distribution of examples in the dataset. In practice, $p(y_C)$ can be approximated as:

$$\hat{p}(y_C) = \frac{1}{N}\sum_{n=1}^{N} p(y_C|x_n) \tag{5.5}$$

The IS considers two qualities that generated examples should simultaneously verify. First, each generated example x_n should be clearly associated

to a class, *i.e.* the classifier output $p(y_C|x_n)$ should have low entropy with high probability for one class and low probability for others. This is partly a measure of the realism of generated examples, although its interpretability as such depends on the confidence of the classifier model on real examples. Second, the model should generate diverse examples in terms of associated classes, *i.e.* $p(y_C)$ should have high entropy, ideally with equal probability for each class. A well-performing generative model according to these criteria will thus yield a higher IS.

In the current study, examples in the DCASE2017 dataset are annotated with one of 15 ambiance classes (see Section 5.3.2). Contrary to synthesis tasks conditioned on class labels, for which the IS is initially proposed, the diversity of generated samples is already ensured by strong conditioning on input third-octave representations. However, the IS still provides information about the presence of sufficient fine-grain information in generated examples to identify types of sound environments. Furthermore, the DCASE2017 evaluation set is balanced in terms of classes. Thus, a lower entropy of the marginal class probability $p(y_C)$ for generated samples would likely indicate limitations of the synthesis model in generating examples for some ambiances, and result in lower IS.

The DCASE2017 task 1 baseline model is taken as a sound environment classifier [56]. It is re-implemented identically for the present study, and the implementation is validated against published performances. This classifier is a multi-layer perceptron (MLP) architecture with 3 fully connected layers. Input representations are log-Mel spectrograms computed on frames of 40 ms with 50% overlap. The Mel filterbank includes 40 bands in the $[0Hz - 22050Hz]$ range. Individual input examples are textures of 5 frames reshaped into a vector of dimension 200×1. The first two fully connected layers have 50 output neurons and are followed by Rectified Linear Unit (ReLU) activations. Furthermore, dropout is applied after these layers with probability 0.2. The last layer outputs a vector of dimension 15×1 and is followed by a softmax activation. Each output value then corresponds to the predicted probability that the input example is associated to each of the 15 ambiance classes. The model is trained using the Adam gradient descent algorithm with a learning rate of 0.001, for 200 epochs or until convergence is observed on the validation subset. The model is trained once on the development set of the DCASE2017 dataset (see Section 5.3.3 for details), and predicts class distributions for real and generated samples of the evaluation set. However, because in the considered task setting models only reconstruct spectral information in the $[20Hz - 8000Hz]$ range, training examples are pre-processed by a band-pass filter with these cutoff frequencies.

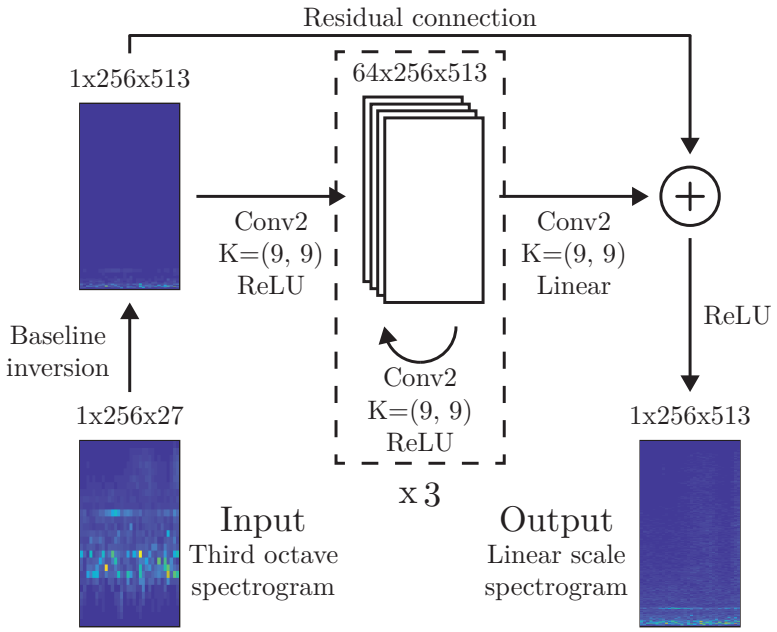


Figure 5.4: Proposed architecture to refine estimations of fine-band spectrograms obtained by application of the third-octave transform pseudoinverse. A convolutional neural network outputs corrections added to the input estimation to produce the prediction.

5.4 Deterministic approach

5.4.1 Architecture

As discussed in Section 5.3.3, applying the pseudoinverse of the matrix associated with the third-octave transform filterbank provides optimal reconstruction with no a priori information on the audio data. The first proposed approach, termed *CNN* in the remainder of this chapter, thus consists in extracting such knowledge from the training dataset, to improve on initial estimations of the fine-band spectrogram obtained by applying the transform pseudoinverse. To do so, a deterministic convolutional neural network is considered in the setting illustrated in Figure 5.4. The pseudoinverse transformation associated to the third-octave filterbank is first applied to an input 3 s third-octave spectrogram with 26 frequency bands in the $[20Hz-8000Hz]$ range. This initial estimation of the fine-band spectrogram (513 bands) is input to a convolutional neural network, and further added to the model's output in a residual connection. Thus, during optimization the output of the

Table 5.1: Evaluation metrics for examples synthesized with the proposed CNN compared to the pseudoinverse baseline, in terms of reconstruction error (resp. signal to reconstruction ratio, log-spectral distance, perceptual similarity metric) and inception score. Statistics are computed on the DCASE2017 evaluation dataset (n=1620). Results shown in bold for specific metrics are not statistically different from the best performing system ($p<0.05$).

Model	SRR Oracle (dB)	SRR LWS (dB)	LSD	PSMt	IS
Reference	$+\infty$	-3.94 ±0.41	0	0.933±0.048	0.089
Pseudoinverse	**18.01 ±7.12**	-3.95 ±0.37	0.416 ±0.158	0.644 ±0.045	0.086
CNN	**18.11 ±7.76**	-3.69 ±0.51	**0.358 ±0.126**	0.695 ±0.044	**0.089**

last convolutional layer is compared to the error between the ground truth fine-band spectrogram and the initial pseudoinverse estimation. In other terms, the convolutional architecture does not predict absolute magnitude values of the spectrogram, but small correction terms reducing errors in the pseudoinverse estimation.

The neural architecture is composed of 5 convolutional layers with kernel size $K = (9, 9)$. No filter stride or downsampling layers are applied, and the input of each layer is zero-padded so that the input and output representations have identical dimensions in time and frequency. The number of convolution channels (64) in hidden layers is also kept constant throughout the network, and the output representation has a single channel. Rectified Linear Unit (ReLU) activations are applied to the output of each layer except the last, as the network should predict zero-centered corrections of the pseudoinverse estimation error. The network is characterized by a total of 1.2 million parameters.

The loss minimized during training is the L_1 distance between pseudoinverse estimations refined by the convolutional network outputs $\hat{Y}_n$ and ground truth fine-band spectra Y_n, averaged over N_b examples in batches:

$$L(\hat{Y}_n, Y_n) = \sum_{n=1}^{N_b} \sum_{t,f} |\hat{Y}_n(t, f) - Y_n(t, f)| \tag{5.6}$$

The model is trained using the Adam gradient descent algorithm with a learning rate of 0.0001 on batches of 32 examples for 50 epochs.

5.4.2 Evaluation

The performance of the proposed convolutional network is evaluated in comparison to the pseudoinverse baseline on metrics presented in Section 5.3.5. Table 5.1 summarizes the mean and standard deviations of evaluation metrics over the DCASE2017 evaluation dataset (n=1620). All metrics are computed on synthesized audio where the phase is recovered using the LWS algorithm unless specified. The SRR as described in eq.5.2 involves a sample-wise difference between the reconstructed and ground truth waveform signals. Thus, it is very sensitive to small variations in the phase of synthesized audio. The LWS algorithm estimates a phase component through an iterative optimization process, and its output thus varies from the ground truth phase. As a result, although the SRR of audio with phase recovered through the LWS algorithm is low for all approaches, it is similarly low for reference samples where the phase component is discarded then retrieved with the LWS method. This metric is thus not discussed further.

Computing the SRR on waveform audio synthesized with the oracle (ground truth) phase results in similar performances of the proposed CNN and the pseudoinverse baseline. However, the CNN significantly improves estimations in terms of the LSD. The inception score for the CNN is close to that of reference sound scenes (0.89). This indicates that the spectral properties necessary for the DCASE2017 baseline classifier to identify the type of sound environment are sufficiently well recovered by the approach. However, informal listening shows that the fine structure in generated spectra is not well recovered, which results in a difficulty to identify harmonic sources. This is illustrated by examples of synthesized spectrograms in Figure 5.5. Compared to the pseudoinverse baseline (b), the CNN improves estimations mostly by smoothing discontinuities at third-octave filter cutoff frequencies, but is unable to generate fine harmonic structures found in the reference spectrogram.

Several factors may limit the ability of the proposed convolutional network to recover fine-grain structure in generated spectra. As a deterministic architecture, it may learn to account for overall properties of spectra associated to environmental sounds in the training dataset, but it is much more challenging to generate specific local variations that are characterizing different sources in the scenes. Furthermore, only using error-based regression losses (e.g. L_1) shifts the focus of the training process towards regions where the estimation error is large. As discussed in Section 5.3.3 (Figure 5.3(b)), the estimation obtained by applying the pseudoinverse of the third-octave filterbank yields, on average, a "sawtooth"-like error of fine-band magnitude

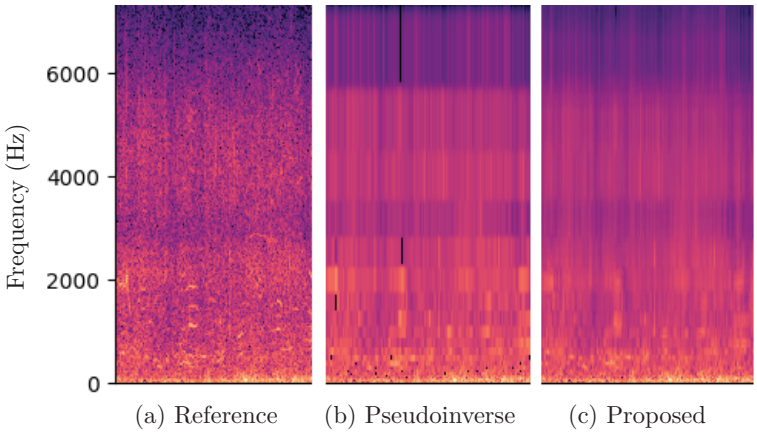


Figure 5.5: Example of spectrograms generated with the pseudoinverse baseline (b) and the proposed CNN (c). The CNN corrects discontinuities at third-octave filter cutoff frequencies, but does not produce fine structure.

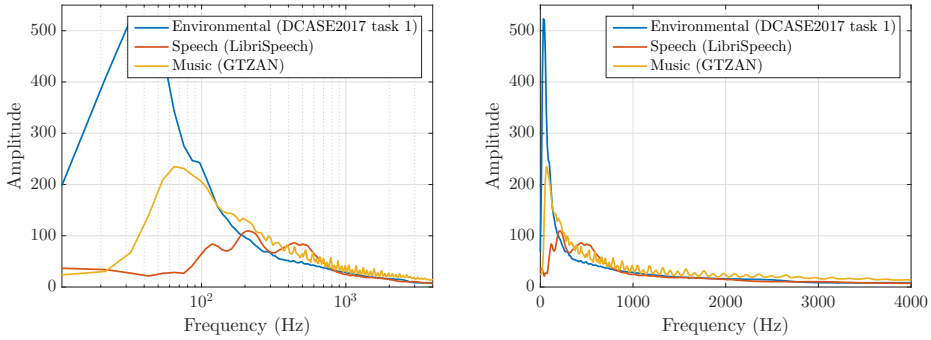


Figure 5.6: Average spectra of normalized audio extracts in datasets of environmental sounds (DCASE2017 task 1), clean speech (LibriSpeech) and music (GTZAN).

estimations. The convolutional architecture must also compensate for this error term, which is generally of higher amplitude than fine-grain variations in magnitude lost during third-octave analysis. Because of this discrepancy, the network is likely to prioritize the matching of the overall spectral shape while disregarding local variations, likely to convey the spectral attributes necessary to trigger human recognition of sound objects.

Table 5.2: Signal-to-reconstruction ratio obtained by the proposed CNN in comparison to the pseudoinverse baseline, as a function of sound environment types in the DCASE2017 evaluation dataset (n=108).

	SRR, oracle (dB)	
Ambiance	Pseudoinverse	CNN
beach	10.07 ± 3.60	9.77 ± 3.59
bus	26.49 ± 6.98	27.33 ± 7.21
cafe/restaurant	12.76 ± 3.13	12.95 ± 3.12
car	28.25 ± 5.77	30.57 ± 7.05
city center	18.74 ± 2.78	18.11 ± 2.78
forest path	18.05 ± 5.23	17.12 ± 5.35
grocery store	17.78 ± 2.73	17.71 ± 2.91
home	11.16 ± 4.60	10.26 ± 3.91
library	18.45 ± 6.58	18.27 ± 6.43
metro station	15.46 ± 3.98	15.55 ± 4.20
office	11.75 ± 2.42	10.77 ± 2.20
park	16.57 ± 3.37	16.55 ± 3.19
residential area	17.77 ± 5.14	17.27 ± 5.31
train	26.71 ± 4.04	27.84 ± 3.92
tram	20.06 ± 5.58	21.65 ± 5.60

The properties of environmental sounds further reinforce limitations of fine structure generation in deterministic models trained on regression loss functions. Environmental sounds are generally characterized by full-band spectra with polynomial energy decay as a function of increasing frequency. Such properties are found in all types of sounds, but induce particularly high in environmental sounds where energy is found in very low frequencies (20 Hz-50 Hz). This is illustrated in Figure 5.6 that compares the average normalized spectra of environmental sounds, clean speech, and music, computed on recordings in the DCASE2017 task 1 [56], LibriSpeech [68], and GTZAN [1] datasets respectively. This issue has to be faced for each computational approximation paradigm. For example, it has been tackled for the non negative matrix factorisation techniques by considering scale invariant divergence such as the Itakura-Saito divergence [128].

Averaging along frequency bands in regression loss functions thus gives the most importance to low-frequency content, whereas estimation errors in high frequencies are likely ignored. This behavior appears in the CNN, and is

[1] http://marsyas.info/downloads/datasets.html

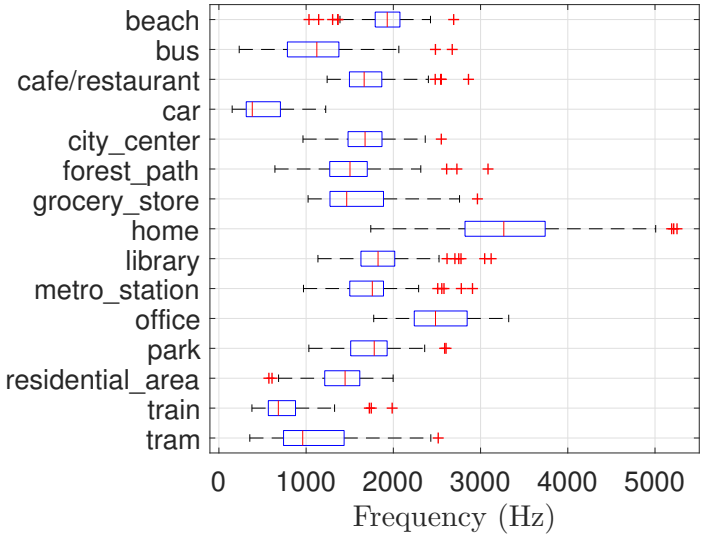


Figure 5.7: Box plots of the distributions of spectral centroids in sound scenes of the DCASE2017 dataset, separated by labels of type of sound environment (n=108).

demonstrated by computing reconstruction metrics on extracts representing separate ambiances. Results are presented in Table 5.2, and Figure 5.7 shows the distribution of spectral centroids in scenes associated to each ambiance for reference. The proposed model yields improvements on the SRR for ambiances with primarily low-frequency content, such as *car* or *tram*. Its performance is however similar to the pseudoinverse for calmer environments where low-frequency information is less important, for example *home* and *office*.

A potential solution to this problem is to apply a "whitening" transformation or a logarithm function to magnitude spectrograms before computing losses. This scales amplitudes to the same order of magnitude across frequencies, and thus ensures similar contributions from errors in all frequency bands to the global loss (scale invariance). However, these methods are found difficult to control as they also amplify high-frequency noise, which result in unstable training procedures.

5.5 Generative approach

Results discussed in Section 5.4.2 underline two main considerations for the design of deep learning spectral feature inversion models. First, producing spectra with fine structures is difficult with fully deterministic architectures. The role of these structures is to enable human identification of content within synthesized scenes, and perfect reconstruction is not required to this aim. Stochastic generative approaches have the potential to improve the quality of generated examples on this aspect, by introducing more degrees of freedom in the synthesis process. Second, even if the model is capable of producing small scale structures, the constraint of correspondence between generated and associated reference spectrograms must be enforced in the considered feature inversion problem. Doing so with regression losses (e.g. L_1 distance) can result in high-frequency content being ignored during the training process. This section introduces a new stochastic generative model with an architecture specifically designed to synthesize fine structure with equal focus across all frequencies.

5.5.1 Generative adversarial networks

The current task of synthesizing short-term Fourier transform magnitudes from log-frequency spectra can be formulated as the upsampling (or interpolation) of third-octave spectrograms in the frequency dimension. This is a well-known task in the image processing community, where state-of-the-art performances are achieved by generative adversarial networks [129, 130, 131]. Similar systems can be applied to the spectral feature inversion task. This section thus presents the general motivations and design paradigms of adversarial approaches to solving synthesis tasks.

Generative adversarial networks are first proposed in [132] with the general framework shown in Figure 5.8. GANs are composed of two neural architectures: a generator and a discriminator. In an unconditioned setting, the generator G takes as input a random vector z sampled from a distribution $p(z)$, and outputs a prediction $\tilde{x} = G(z)$. The discriminator D is then tasked to identify fake samples from the generator and real samples $x \sim \mathbb{P}_x$, where the discriminator learns to infer $\mathbb{P}_x$ from examples in the training dataset. The discriminator outputs a scalar value in the $[0, 1]$ range, where 0 corresponds to a fake sample and 1 corresponds to a real sample. Using a cross-entropy loss, the associated objective function of the optimization process is a minimax objective:

$$\min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_x}\left[log(D(x))\right] + \mathbb{E}_{\tilde{x}\sim\mathbb{P}_{\tilde{x}}}\left[log(1 - D(\tilde{x}))\right] \tag{5.7}$$

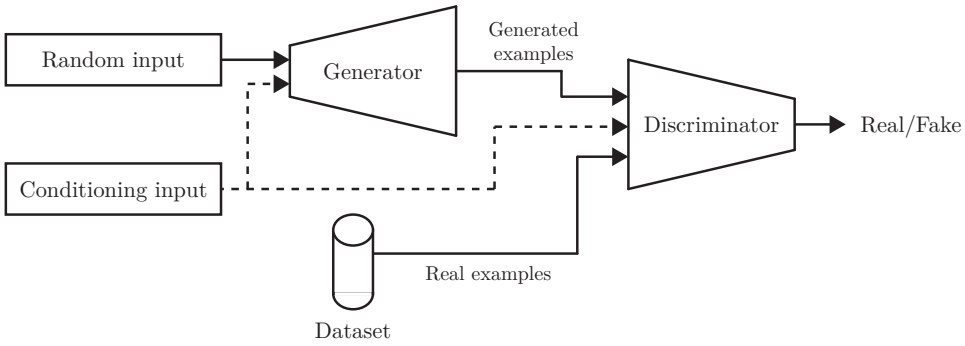


Figure 5.8: Original framework of generative adversarial networks. A generator architecture synthesizes examples from input random vectors. A discriminator is tasked to identify generated examples from real examples in the dataset, and both models are trained jointly in a minimax setting. An additional input optionally conditions synthesis.

In other terms, the generator is trained to match the distribution $\mathbb{P}_x$ so that the discriminator cannot identify fake samples from real samples.

The authors of [133] note that the original formulation of the adversarial loss in eq. 5.7 leads to unstable training in some cases. If the discriminator is trained to optimality, the gradient of the cross-entropy loss is close to zero. Backpropagating this gradient then yields very small updates to the generator parameters, which stops the convergence towards an acceptable solution. Conversely, an insufficiently trained discriminator provides no useful gradients for the generator to shift its output distribution towards that of the dataset $\mathbb{P}_x$. As an alternative, they propose the Wasserstein-GAN framework, in which the discriminator is replaced by a critic architecture that can be trained to optimality while retaining useful gradients. Instead of predicting a label of fake or real samples, the critic outputs a real-valued score. During critic training the score is maximized for real samples and minimized for fake samples from the generator. The generator is then trained to maximize the critic score associated to its output samples:

$$\min_D \mathbb{E}_{\tilde{x}\sim\mathbb{P}_{\tilde{x}}}\left[D(\tilde{x})\right] - \mathbb{E}_{x\sim\mathbb{P}_x}\left[D(x)\right] \tag{5.8}$$

$$\min_G \; -\mathbb{E}_{\tilde{x}\sim\mathbb{P}_{\tilde{x}}}\left[D(\tilde{x})\right] \tag{5.9}$$

The authors show that the critic approximates the Wasserstein-1 (Earth Mover) distance if its associated function is K-Lipschitz. This constraint is

initially achieved by clipping the critic parameters to $[-c, c]$ after each critic optimization step, where c is a small constant (e.g. 0.01). In [134], task examples are found where weight clipping leads to failure of convergence of the training process, or where the generator fails to capture the underlying distribution $\mathbb{P}_x$ of real examples. Instead, a gradient penalty is proposed to enforce the Lipschitz constraint. The gradient penalty is added as a regularization term to the loss minimized during critic optimization (see Section 5.5.4 for details).

Following this argument, the current study addresses the design of a generative model trained in a Wasserstein-GAN setting with gradient penalty to reconstruct short-term Fourier transform spectrograms by upsampling the input log-frequency spectral representation.

5.5.2 Generator architecture

The structure of the proposed generator model is shown in Figure 5.9. For a third octave filterbank with N_b bands, the generator is a group of N_b convolutional networks, each tasked with producing the fine-band magnitude estimations in the frequency range of one third-octave band. Each subnetwork takes the full third octave spectrogram of dimension 256×26 as input, and outputs a matrix of dimension 256×513 that corrects the full-band pseudoinverse spectrogram estimation.

In the Fourier domain, the third-octave filterbank Φ (see Figure 5.2(a)) verifies:

$$\sum_{b=1}^{N_b} = 1 \tag{5.10}$$

within the considered frequency range of $[20Hz, 8000Hz]$, *i.e.* the filterbank conserves energy in that range. Thus, the output estimation $\tilde{x}$ can simply be obtained from subnetwork contributions $\tilde{x}_b$ as

$$\tilde{x}(t, f) = \sum_{b=1}^{N_b} \Phi_b(f) \cdot \tilde{x}_b(t, f) \tag{5.11}$$

where Φ_b is the frequency response of the third-octave filter at band b. During backpropagation, the gradient of the loss function L with respect to subnetworks outputs $\tilde{x}_b$ is:

$$\frac{\partial L}{\partial \tilde{x}_b} = \frac{\partial L}{\partial \tilde{x}} \frac{\partial \tilde{x}}{\partial \tilde{x}_b} = \frac{\partial L}{\partial \tilde{x}} \cdot \Phi_b \tag{5.12}$$

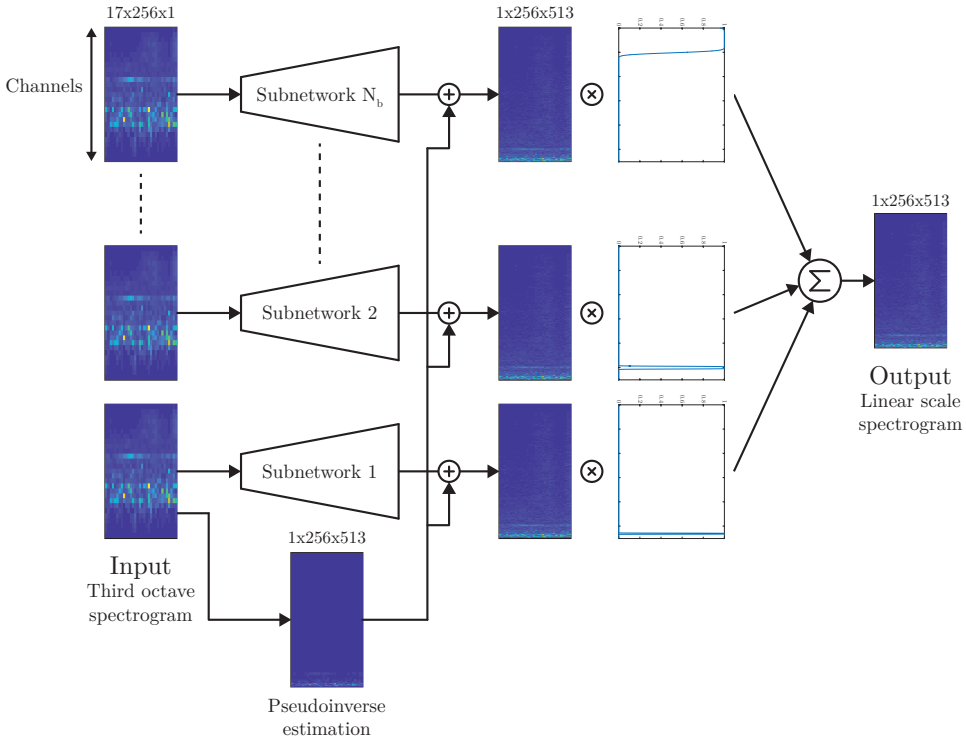


Figure 5.9: Proposed generator approach, where separate deep neural networks estimate parts of the fine-band magnitude spectrogram corresponding to each third-octave filter. Combining contributions from the subnetworks produces the final estimation, and ensures that gradients in subnetworks are not affected by large magnitude differences across frequencies.

Because the frequency response of third octave filters quickly drops to zero outside of their bandwidth (see Figure 5.2(a)), gradients of errors in fine-band frequency bins outside this range are multiplied by 0 during backpropagation. As a result, the prediction error of any given frequency bin affects the gradient in a maximum of two subnetworks corresponding to adjacent bands. Thus, issues observed in Section 5.4.2, where prediction errors in high frequencies are ignored due to larger contributions to the gradient of prediction errors in low frequencies, do not appear in this formulation. This allows us to potentially retain useful gradients across all frequencies when quantifying errors with loss functions that average errors over frequencies.

The architecture of convolutional subnetworks is shown in Figure 5.11.

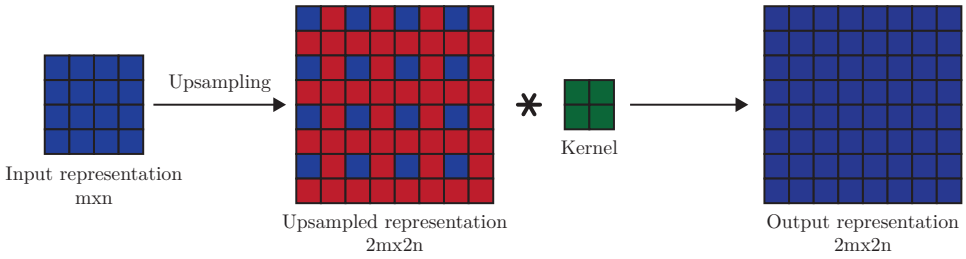


Figure 5.10: Example of transposed convolution layer that performs the parametric upsampling of an input representation, here with an upsampling factor (fractional stride) of 2 and kernel size 2×2

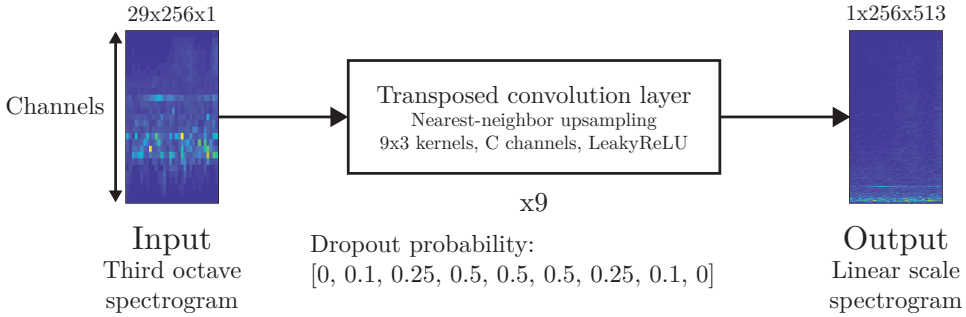


Figure 5.11: Architecture of individual subnetworks in the generator model. 9 transposed convolution layers upsample the input third-octave spectrogram into an estimation of the fine-band spectrogram. A random component is introduced with dropout during both learning and evaluation. The number of channels C in hidden layers is a function of the number of frequency bins associated with the third-octave filter inverted by the network.

Subnetworks are composed of 9 transposed convolution layers that each upsample the input representation along the second (frequency) dimension by a factor of two. The principle of transposed convolutions, also known as fractionally-strided convolutions, is illustrated in Figure 5.10. In this example, the input matrix of dimension $m \times n$ is first upsampled by a factor of 2 by inserting zeros. The upsampled matrix is then zero-padded and filtered by kernels, similarly to regular convolution layers. The output is thus a matrix of dimension $2m \times 2n$. In the proposed model, only the second (frequency) dimension is upsampled and the first (time) dimension remains unchanged throughout the network. Furthermore, nearest-neighbor upsampling is pre-

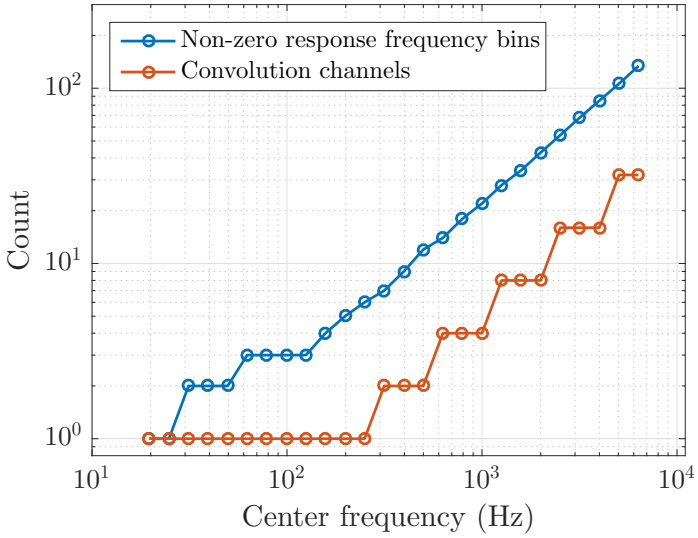


Figure 5.12: Number of frequency bins in the bandwidth of third-octave filters and number of channels allocated to corresponding subnetworks in the generator architecture. Both quantities are doubled with each octave.

ferred to zero-insertion to alleviate "checkerboard" patterns in output representations. Third-octave bands in input spectrograms are viewed as separate channels by the network, and the input dimensions are $26 \times 256 \times 1$ (resp. channels, time, frequency). Thus, the network progressively upsamples the representation from 1 to 513 bands. The motivation of this approach is that the frequency dimension of third-octave spectra is logarithmically downsampled from that of fine-band spectra. Transposed convolution layers perform linear-step upsampling, thus they should not upsample the frequency dimension of third-octave spectra directly. The kernel size of filters is 9×3 (resp. time, frequency), and is identical for all convolution layers.

Each subnetwork predicts short-term Fourier transform magnitudes within the bandwidth of a different third-octave filter, which corresponds to $a \cdot 2^{b/3}$ frequency bands where a depends on the precision of the spectrogram. Because the resulting task difficulty increases with the center frequency of logarithmic bands, the capacity of corresponding subnetworks should vary accordingly. Here, a low number of channels $C = 2$ is allocated to subnetworks associated with the lowest third-octave bands. Every three bands this number of channels is doubled, up to $C = 32$ in the subnetwork corresponding to

the highest third-octave band. However, the number of channels only starts increasing from the third-octave band centered at 160 Hz, in order to reduce the total number of parameters in the generator. Figure 5.12 summarizes the allocated number of channels compared to the number of fine bands that each subnetwork is tasked to predict, *i.e.* where the frequency response of the corresponding third-octave filter is not zero. The number of channels is always set as a power of two to maximize the benefits of distributed parallel computing, which minimizes the training time.

After each hidden layer, a Leaky Rectified Linear Unit (LeakyReLU) with slope 0.01 is applied to the representation. Because the generated spectra should be non negative, a Rectified Linear Unit (ReLU) activation is applied to the output spectrogram after combining subnetwork contributions. Lastly, to train the generator in an adversarial setting, a random component should be introduced. The authors of [135] note that when the generator is strongly conditioned, for example with a spectral representation, using a random vector as an additional input does not always result in diverse outputs. In some cases the generator may choose to ignore this additional input and act as a deterministic architecture. Instead, they propose to force stochastic predictions by applying a dropout layer after some of the convolutional layers. Dropout layers set each element of the input representation to 0 with probability p, then scales the remaining values by $\frac{1}{1-p}$, where p is a fixed hyperparameter of the model. In the literature, dropout is typically applied only during training of models as a regularization method to reduce overfitting, and is replaced by the identity function at evaluation. In [115] dropout layers are retained at evaluation as a source of randomness. We believe that considering those layers both at training and evaluation is beneficial in strongly conditioned synthesis tasks. However, applying dropout in early layers where the frequency dimension is low may result in significant loss of information. Furthermore, applying dropout on the last layer leads to zeros in the output spectrogram. Thus, dropout probabilities for the 9 layers of each subnetwork are set to $[0, 0.1, 0.25, 0.5, 0.5, 0.5, 0.25, 0.1, 0]$. In total, the generator architecture contains about 0.8 million parameters.

5.5.3 Critic architecture

In typical Wasserstein-GAN settings, critic architectures output a real-valued score that evaluates the fidelity of the entire input representation. This critic input can be the output of the generator with distribution $\tilde{x} \sim \mathbb{P}_g$, or an item of the dataset of real examples with distribution $x \sim \mathbb{P}_r$. In the current study, examples consist in about 3 s of audio data (256 frames of 46.4 ms with 75%

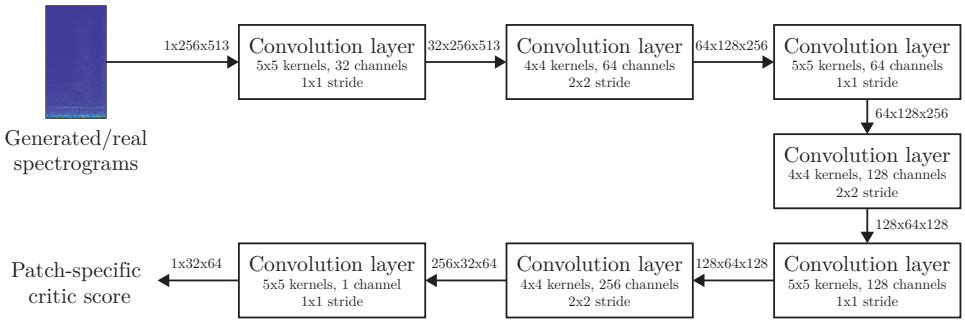


Figure 5.13: Architecture of the critic model rating the realism of patches in generated and real fine-band spectrograms.

overlap). Instead of the full example, the proposed critic architecture rates the realism of time-frequency patches in input spectrograms. This allows the critic model to focus on the fidelity of reconstructed patterns at a smaller scale, rather than the overall likeness between generator outputs and real examples from the dataset. The proposed architecture is shown in Figure 5.13. It is composed of 7 convolution layers each followed by LeakyReLU nonlinear activations. The third, fifth, and seventh layers are strided convolutions that downsample the representation by a factor of 2 in the time and frequency dimensions. The kernel size in these layers is 4×4, whereas in other layers a filter size of 5×5 retains equal dimensions in input and output representations. The number of channels is set to 32 in the first layer, and doubles in each strided convolution layer up to 256 in the penultimate layer. This compensates the dimensionality reduction yielded by strided convolutions, to retain sufficient information capacity throughout the network. The final layer outputs a single channel corresponding to critic scores. The output is a matrix of dimension 32×64 where values correspond to a small overlapping patches of the input fine-band spectrogram. The dimensions of these patches is given by the total receptive field of stacked filters in the model. Each value in the output matrix follows the same principles as scalar critic outputs in typical approaches: the discriminator is trained so that high positive values are output for real examples from the dataset, and high negative values are output for fake examples from the generator. The total number of parameters in the critic architecture is of the order of 1 million.

5.5.4 Training process

The generator and critic architecture are jointly trained by the Wasserstein-GAN algorithm with gradient penalty [134]. For generator outputs $\tilde{x} \sim \mathbb{P}_g$ and real examples $x_n \sim \mathbb{P}_x$, the critic (D) and generator (G) losses (resp. L_D, L_G) minimized during training are:

$$L_D = \frac{1}{N}\sum_{n=1}^{N} D(\tilde{x}) - D(x) + \lambda \left(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1\right)^2 \tag{5.13}$$

$$L_G = \frac{1}{N}\sum_{n=1}^{N} -D(\tilde{x}) \tag{5.14}$$

where N is the batch size, $\nabla_{\hat{x}}$ denotes the gradient with respect to $\hat{x}$, and λ is the gradient penalty coefficient set to 10 in this study following [134]. The $\hat{x}$ term in the gradient penalty corresponds to the weighted sum of a real sample x and a generated sample $\tilde{x}$:

$$\hat{x} = \varepsilon x + (1 - \varepsilon)\tilde{x} \tag{5.15}$$

where ε is a random scalar drawn from a $[0, 1]$ uniform distribution at each iteration.

In addition to the realism obtained by training the generator in an adversarial setting, generated spectrograms should correspond in content to the input third-octave measurements. Following [115], this is achieved by introducing an additional loss term in the generator optimization. The L_1 loss comparing generator outputs $\tilde{x}_n$ to corresponding ground truth fine-band spectrograms x_n is taken:

$$L_G = \frac{1}{N}\sum_{n=1}^{N} [-D(\tilde{x}) + \alpha|\tilde{x}_n - x_n|] \tag{5.16}$$

where α is a scalar coefficient determining the trade-off between the adversarial and reconstruction components of the loss function. $\alpha = 100$ results in the same order of magnitude for the two components in this study, and is thus taken in reported experiments.

In the Wasserstein-GAN formulation, the generator benefits from training the critic to optimality (see Section 5.5.1). Thus, the critic architecture is initially trained for 10000 iterations using examples from the dataset as real samples, and outputs of the generator with randomly initialized parameters as fake samples. The Adam gradient descent algorithm computes parameter

Table 5.3: Performance of the proposed Ad-SBSR generative approach compared to the convolutional network (CNN) in Section 5.4, pseudoinverse estimations, and the *AdVoc* generative baseline. Metrics are evaluated on the DCASE2017 evaluation dataset (n=1620). Results shown in bold for specific metrics are not statistically different from the best performing system ($p<0.05$).

Model	SRR Oracle (dB)	LSD	PSMt (LWS)	IS
Reference	$+\infty$	0	0.933±0.048	0.089
Pseudoinverse	**18.01 ±7.12**	0.416 ±0.158	0.644 ±0.045	0.086
CNN	**18.11 ±7.76**	**0.358 ±0.126**	0.695 ±0.044	0.089
AdVoc	14.89 ±6.51	0.487 ±0.175	**0.739 ±0.051**	0.083
Ad-SBSR	**17.71 ±7.24**	0.407 ±0.142	0.656±0.048	**0.092**

updates with a learning rate of 0.0001 over batches of 16 examples. Once the pre-trained critic is obtained, both architectures are jointly trained by alternating generator and critic optimization iterations. In a group of 6 subsequent iterations, each with a separate batch of examples, 5 iterations are dedicated to the optimization of the critic parameters, and one to the optimization of the generator parameters. This allows the critic to remain near-optimal whenever an optimization step of the generator is computed. The model is trained for a total of 50 epochs.

5.5.5 Evaluation

The proposed generative approach, named Adversarial Spectral Band-Specific Reconstruction (Ad-SBSR), is compared to the CNN deterministic approach of Section 5.4, as well as the pseudoinverse reconstructions and *AdVoc* deep generative baselines. Table 5.3 shows reconstruction metrics and inception scores computed for all methods on the DCASE2017 evaluation dataset (n=1620).

The deterministic CNN is trained on the L_1 distance, and thus strictly focuses on matching ground truth spectral magnitudes. This results in high performance on signal processing reconstruction metrics (SRR, LSD). On the contrary, the *AdVoc* generative baseline achieves worse performance on these reconstruction metrics, but improves synthesis quality in terms of the perceptual similarity metric (PSMt). This is attributed to the combination of loss functions minimized in its training process (see Section 5.3.3). During the training of *AdVoc* the L_1 is balanced by the additional adversarial loss, which favors the correspondence between distributions of generated and real

spectrograms regardless of the reference associated to synthesized examples, in order to allow the system to generate more realistic samples. However, this improvement is only evidenced with the PSMt computed on waveform signals with LWS phase reconstruction. Synthesizing samples with oracle phase yields PSMt correlated to the LSD metric. The LWS thus converges to better phase component estimates on spectra produced by *AdVoc* than by other investigated methods.

Short-term Fourier transform spectrograms generated by the Ad-SBSR architecture obtain similar SRR compared to examples generated by the pseudoinverse and CNN. We hypothesize that this is due to the addition of pseudoinverse estimations to the output of band-specific generator subnetworks, as it provides a strong initial point that the model is trained to improve. In contrast, *AdVoc* is only given the pseudoinverse estimations as inputs, and must fully reconstruct STFT magnitudes. However, enforcing this strong constraint on the Ad-SBSR architecture output may ultimately be detrimental to the ability of the training process to converge towards a perceptually more relevant solution: because the pseudoinverse already provides a good reconstruction in the least-squares sense, the initial gradient of the L_1 regression loss term is expected to be small, *i.e.* examples may not significantly diverge from the pseudoinverse solution during optimization. Still, the Ad-SBSR architecture improves the inception score over other approaches, indicating an ability to generate properties necessary to the identification of sound environment types by the DCASE2017 classifier.

Quantitative results discussed here are encouraging and demonstrate the potential of the Ad-SBSR to tackle the long standing issue of frequency axis magnitude assymetry in spectral audio processing. That being said, some experiments remain to be conducted to provide definitive evidence and hopefully obtain production-ready models supported by relevant evaluation metrics. Design parameters of the Ad-SBSR approach should be further explored, including an ablation study to assess performance gains associated to band-specific reconstruction and stochastic adversarial training respectively. Because objective metrics provide limited information on the perceptual quality of synthesized sounds, subjective evaluation through listening test is also necessary as a conclusive indicator of performance.

Conclusion

This thesis focused on proposing tools to infer humanly interpretable information about urban sound environments measured by large-scale acoustic sensor networks. Specifically, two contributions were proposed to investigate the modeling potential of deep learning approaches for those matters: the prediction of perceptual attributes and the synthesis of sound scenes from privacy-aware spectral representations.

Despite its high potential impact on the characterization of sound environments, the task of predicting perceptual attributes remains relatively unexplored by the Detection and Classification of Acoustic Scenes and Events (DCASE) community. This area of research is inherently interdisciplinary, at the frontier of machine learning and psychoacoustics. As such, the task of predicting perceptual quantities related to the activity of sources of interest, on which attributes of soundscape quality depend, differs from typically considered event detection contexts. Variations in perceptual descriptors occur on longer time scales, and sources composing polyphonic sound environments interact more strongly through salience.

These properties result in difficulties to obtain relevant labels in large datasets of sound scenes necessary to the training process of deep learning architectures. In this context, the use of controlled corpora is proposed in Chapter 2. This approach provides the ability to develop effective methods for the automatic annotation of perceived source activity. Through data augmentation at the scene level, sound scene simulation tools then allow us to generate arbitrary large datasets at a low human labor cost, with sufficient diversity and annotation quality to train deep learning architectures on the desired task.

Accurately predicting perceived source presence labels is achieved with convolutional and recurrent deep learning models, as described in Chapter 3. However, the generalization to target sound environments highly depends on the localization properties of simulated corpora. This issue is addressed to some extent by using on-site recordings of sound sources in the simulation

process of sound scenes. Chapter 4 further develops an approach that exploits the availability of very large unlabeled datasets with an original use of transfer learning techniques. Useful knowledge about the target domain is extracted by learning latent audio representations in unsupervised or semi-supervised settings. Results show that considering these representations as an initial estimation of deep learning model parameters contributes to the domain specialization of trained predictive architectures when the supervised target task is trained on simulated data with limited content correspondence, or when the quantity or diversity of localized simulated data is reduced. In particular, these results show that recording very few clean samples of source occurrences is sufficient in future monitoring projects where large amounts of sensor data are available.

Although the proposed method yields accurate predictions of perceptual quantities on evaluation recordings, we believe that there remains several avenues of research to consider in future work. In the present work, approaches to perceptual source presence annotation and soundscape quality assessment are developed in a simplified context of active listening, in which the passer-by is focused on the sound environment. The annotation of simulated sound scenes in terms of perceived source presence is thus achieved with an emergence indicator that models auditory masking. More complex saliency models are available in the literature that account for the first layers of the auditory system. Such models should be considered in applications where soundscape quality is evaluated from a different perspective, for example sound quality for city residents at home. However, the general methodology for simulating large polyphonic corpora as well as predictive models in Chapters 3 and 4 remains valid for any application scenario.

To maximize the localization characteristics of controlled datasets, simulation parameters should be more fully obtained from target environments. In this study, the considered isolated samples of sources are representative of environments encountered in the application, although the taxonomy of sound sources is somewhat limited. Additional sources may also influence high-level perceptual attributes. We believe that including such sources in datasets and predictions, as well as developing location-specific perceptual models of soundscape quality, would result in a better characterization of sound environments. Furthermore, simulated scenarios are extrapolated from a small corpus of recordings. Manually annotating location-specific recordings in terms of source activity to extract scenario distributions (Section 2.2.2) would thus further improve the adaptability of trained architectures. Alternatively, developing automatic tools for generating large controlled sound scene datasets matching the content and diversity found in

target environments constitutes an important area of research in future work.

The proposed transfer learning approach implemented in this study considers measurements collected over a single week. While these data are sufficient to learn robust latent audio representations, extending the approach with measurements spanning over year-long periods would be beneficial to fully account for the intrinsic variations in sound environments over long periods of time. Similarly, predictive models should be evaluated against *in situ* subjective assessments in diverse conditions, including time of day, day of the week, and seasons.

In the second part of this thesis, the difficult problem of sound synthesis is investigated. Although this area of research is supported by very active deep learning communities, in particular in speech and music applications, there remains several scientific challenges to be addressed. Here, a study is conducted on the synthesis of environmental sound scenes strongly conditioned on log-frequency spectral representations. Specifically, the proposed methods account for the high dissymmetry of spectral audio magnitudes as a function of frequency, which limits the efficiency of cost functions traditionally used in regression tasks. The preliminary study demonstrates the potential of the proposed approach of band specific spectrogram reconstruction in the highly dissymmetric case of environmental sounds.

In the training process of the proposed approach, the cost function implements a trade-off between realism and reconstruction quality constraints in synthesized sound scenes. Realism is evaluated as the correspondence between distributions of generated and ground truth spectrograms by a deep learning critic model, with no clear relation to the perceptual quality of produced examples. Similarly, distance functions quantifying reconstruction errors at evaluation do not necessarily reflect the perceptual differences yielded by prediction errors. To account for these properties, perceptually motivated differentiable loss functions and metrics recently proposed [136] could shift the focus of the training process in synthesis models towards improving the perceptual similarity of reconstructed sounds. Still, the proposed approach contributes to solving the signal processing problems related to spectral characteristics of natural sounds in deep learning paradigms. The discussed results are encouraging, and we hope this contribution will be useful in the very active research domain of sound synthesis, where several emergent architectures continuously improve the state-of-the-art in recent studies [109, 110, 119].

Appendix A

Scene level parameters for acoustic scene simulation

This appendix summarizes the parameters from which the *simScene* sound scene simulation process generates original scenarios used in Section 2.2.2. These parameters are extracted by annotating a corpus of 74 sound scenes recorded in Paris in terms of event and background source activity. In addition to *quiet street*, *noisy street*, *very noisy street* and *park* environment types represented in this corpus, parameters are empirically derived for *square* ambiances with more voice activity. Tables A.1, A.2, A.3, A.4, and A.5 show the respective corresponding values, further used for simulating large datasets in Sections 3.2 and 4.3.1.

Table A.1: Scene level parameters to generate *quiet street* environments.

	Source	Probability of appearance	Event-to-Background ratio (dB)	Time between instances (s)
Background	Traffic	1	-	-
	Voice	0.61	-2.56 ±5.92	-
	Birds	0.18	-4.41 ±6.72	-
Event	Traffic	0.83	7.79 ±4.63	39.97 ±10.41
	Voice	0.93	-4.75 ±1.5	16 ±6
	Birds	0.8	4.7143 ±2.8571	25 ±6

Table A.2: Scene level parameters to generate *noisy street* environments.

	Source	Probability of appearance	Event-to-Background ratio (dB)	Time between instances (s)
Background	Traffic	1	-	-
	Voice	0.75	-3.90 ±2.11	-
	Birds	0.04	-11.63 ±0	-
Event	Traffic	1	3.26 ±3.43	22.60 ±10.28
	Voice	0.87	-7 ±1.5	21 ±6
	Birds	0.59	-0.43 ±2.57	35 ±6

Table A.3: Scene level parameters to generate *very noisy street* environments.

	Source	Probability of appearance	Event-to-Background ratio (dB)	Time between instances (s)
Background	Traffic	1	-	-
	Voice	0.56	-3.92 ±1.02	-
	Birds	0	-	-
Event	Traffic	1.1	2.33 ±2.67	11.47 ±9.24
	Voice	0.53	-8.67 ±1.5	27 ±6
	Birds	0.52	-1.5 ±1.5	40 ±6

Table A.4: Scene level parameters to generate *park* environments.

	Source	Probability of appearance	Event-to-Background ratio (dB)	Time between instances (s)
Background	Traffic	0.6	-	-
	Voice	0.75	2.61 ±3.83	-
	Birds	1	3.00 ±6.60	-
Event	Traffic	0.48	3 ±2.33	45.47 ±6
	Voice	0.9	-6.5 ±1.5	12 ±6
	Birds	1	0 ±2.5	20 ±11.85

Table A.5: Scene level parameters to generate *square* environments.

	Source	Probability of appearance	Event-to-Background ratio (dB)	Time between instances (s)
Background	Traffic	1	-	-
	Voice	1	4.5 ±4	-
	Birds	0.7	-8 ±5	-
Event	Traffic	0.7	4.5 ±3	40 ±7
	Voice	1	-6 ±1.5	10 ±6
	Birds	0.9	3 ±3	22 ±9

Appendix B

Comparison of perceptual responses between recordings and matching synthetic sound scenes

This appendix contains additional material discussed in Section 2.3.1. A principal components analysis is applied to the perceptual assessments obtained during the listening test on 6 recorded and 19 replicated sound scenes. Evaluations of matching recorded and replicated sound scenes by individual participants are projected on the first two components in the resulting space for comparison. Figures B.1, B.2, B.3, B.4, and B.5 show distributions of projections of individual assessments as ellipses, in scenes corresponding to five locations (resp. P3, P4, P8, P15 and P18). The shift in average perceptual evaluation from a recording to the corresponding replicated scene is illustrated by an arrow.

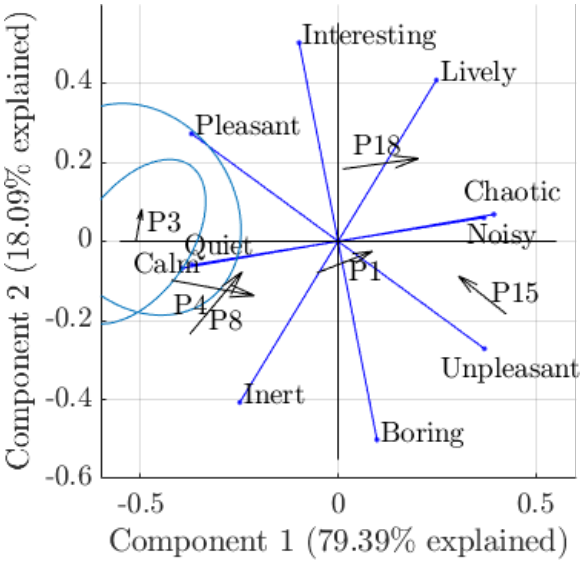


Figure B.1: Biplot of the principal components analysis of average assessments for the 5 high-level perceptual attributes on the 6 recorded and 19 replicated scenes (n=25). Arrows indicate differences between projections of assessments for the recorded (base) and replicated (head) scenes of each location. For the P3 location ellipses show the distributions of individual assessments.

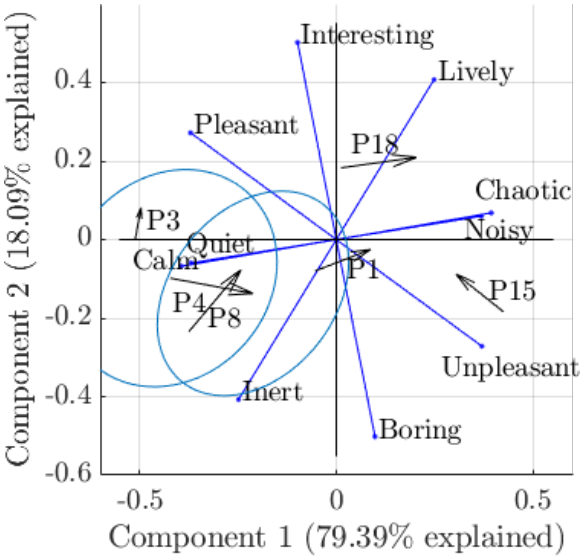


Figure B.2: Biplot of the principal components analysis of average assessments for the 5 high-level perceptual attributes on the 6 recorded and 19 replicated scenes (n=25). Arrows indicate differences between projections of assessments for the recorded (base) and replicated (head) scenes of each location. For the P4 location ellipses show the distributions of individual assessments.

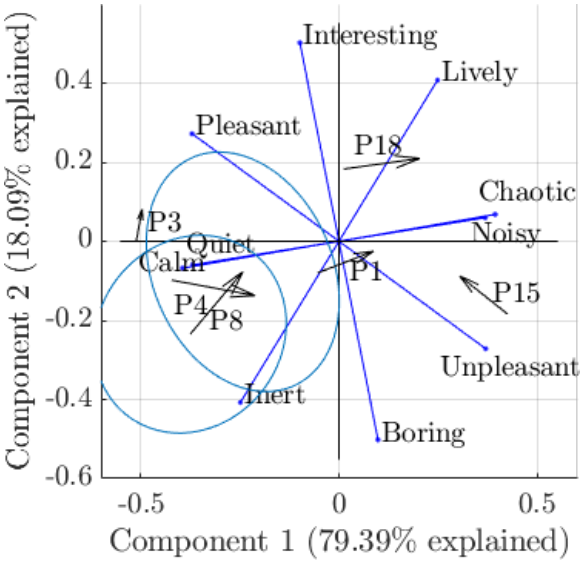


Figure B.3: Biplot of the principal components analysis of average assessments for the 5 high-level perceptual attributes on the 6 recorded and 19 replicated scenes (n=25). Arrows indicate differences between projections of assessments for the recorded (base) and replicated (head) scenes of each location. For the P8 location ellipses show the distributions of individual assessments.

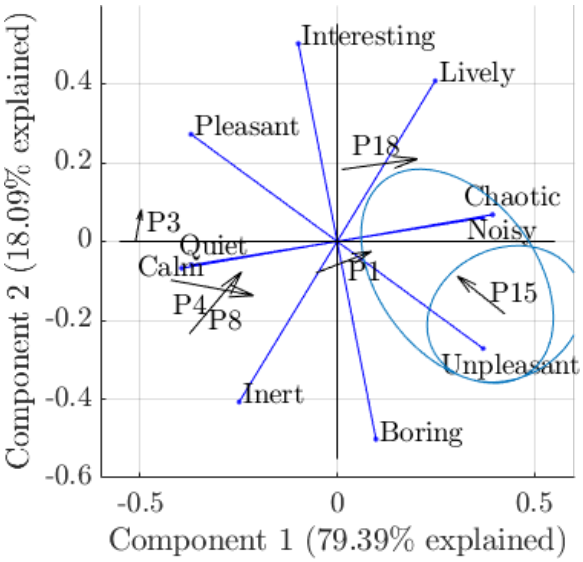


Figure B.4: Biplot of the principal components analysis of average assessments for the 5 high-level perceptual attributes on the 6 recorded and 19 replicated scenes (n=25). Arrows indicate differences between projections of assessments for the recorded (base) and replicated (head) scenes of each location. For the P15 location ellipses show the distributions of individual assessments.

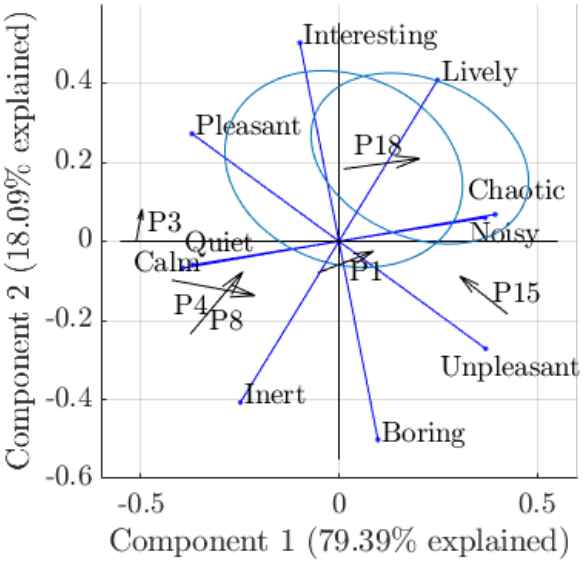


Figure B.5: Biplot of the principal components analysis of average assessments for the 5 high-level perceptual attributes on the 6 recorded and 19 replicated scenes (n=25). Arrows indicate differences between projections of assessments for the recorded (base) and replicated (head) scenes of each location. For the P18 location ellipses show the distributions of individual assessments.

Bibliography

[1] 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.

[2] European environment agency, 'environmental noise'. https://www.eea.europa.eu/airs/2018/environment-and-health/environmental-noise. Accessed: 2020-10-13.

[3] W. Babisch. Road traffic noise and cardiovascular risk. *Noise Health*, 10:27–33, 2008.

[4] S. Pirrera, E. De Valck, and R. Cluydts. Nocturnal road traffic noise: A review on its assessment and consequences on sleep and health. *Environment International*, 36:492–498, 2010.

[5] M. Basner, W. Babisch, A. Davis, M. Brink, C. Clark, S. Janssen, and S. Stanfeld. Auditory and non-auditory effects of noise on health. *The Lancet*, 383:1325–1332, 2014.

[6] EC. 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. *Off. J. Eur. Communities*, 189:12, 2002.

[7] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi. Internet of Things for smart cities. *IEEE Internet of Things Journal*, 1:22–32, 2014.

[8] J. Picaut, A. Can, J. Ardouin, P. Crépeaux, T. Dhorne, D. Écotière, M. Lagrange, C. Lavandier, V. Mallet, C. Mietlicki, and M. Paboeuf. Characterization of urban sound environments using a comprehensive approach combining open data, measurements, and modeling. *J. Ac. Soc. Am.*, 141:3808, 2017.

[9] F. Mietlicki, C. Mietlicki, and M. Sineau. An innovative approach for long-term environmental noise measurement: RUMEUR network in the Paris region. In *Euronoise 2015*, pages 2309–2314, 2015.

[10] X. Sevillano et al. DYNAMAP - development of low cost sensor networks for real time noise mapping. *Noise Mapping*, 3:172–189, 2016.

[11] P. Bellucci, L. Peruzzi, and G. Zambon. LIFE DYNAMAP project: the case study of Rome. *Applied Acoustics*, 117:193–206, 2016.

[12] 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.

[13] B. Pijanowski, L. Villanueva-Rivera, S. Dumyahn, A. Farina, B. Krause, B. Napoletano, S. Gage, and N. Pieretti. Soundscape ecology: the science of sound in the landscape. *BioScience*, 61:203–216, 2011.

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

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

[16] M. Cartwright, J. Cramer, J. Salamon, and J.P. Bello. TriCycle: audio representation learning from sensor network data using self-supervision. In *2019 IEEE Workshop on Applications of Signal Processing on Audio and Acoustics (WASPAA)*, 2019.

[17] A. Cohen-Hadria, M. Cartwright, B. McFee, and J.P. Bello. Voice anonymization in urban sound recordings. In *2019 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.

[18] 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.

[19] R. M. Schafer. *The tuning of the World*. 1977.

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

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

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

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

[24] O. Axelsson, M. Nilssen, and B. Berglund. The Swedish soundscape-quality protocol. *J. Ac. Soc. Am.*, 131:3476, 2012.

[25] 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.

[26] ISO/TS 12913-2:2018. Acoustics - soundscape - part 2: Data collection and reporting requirements. Standard, International Organization for Standardization, Geneva, CH, 2018.

[27] F. Aletta et al. Soundscape assessment: towards a validated translation of perceptual attributes in different languages. In *INTER-NOISE 2020*, 2020.

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

[29] D. Botteldooren, B. De Coensel, and T. De Muer. The temporal structure of urban soundscapes. *Journal Sound Vib.*, 292:105–123, 2006.

[30] 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.

[31] 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.

[32] W. Yang and J. Kang. Acoustic comfort evaluation in urban open public spaces. *Applied Acoustics*, 66:211–229, 2005.

[33] C. Guastavino. The ideal urban soundscape: Investigating the sound quality of French cities. *Acta Acust. unit. Acust.*, 92:945–51, 2006.

[34] 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.

[35] S. Viollon, C. Lavandier, and C. Drake. Influence of visual setting on sound ratings in an urban environment. *Applied Acoustics*, 63:493–511, 2002.

[36] K. Sun. *Audiovisual interaction in the perception and classification of urban soundscapes.* PhD thesis, Ghent University, Faculty of Engineering and Architecture, Ghent, Belgium, 2018.

[37] 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.

[38] O. Axelsson, M. Nilsson, B. Hellstrom, and P. Lunden. A field experiment on the impact of sounds from a jet-and-basin fountain on soundscape quality in an urban park. *Landscape and Urban Planning*, 123:49–60, 2014.

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

[40] J. Lambert, P. Champelovier, and I. Vernet. Annoyance from high speed train noise: a social survey. *J. Sound Vib.*, 193:21–28, 1996.

[41] M. Raimbault, C. Lavandier, and M. Berengier. Ambient sound assessment of urban environments: field studies in two French cities. *Applied Acoustics*, 64:1241–1256, 2003.

[42] A. Fiebig, S. Guidati, and A. Goehrke. Psychoacoustic evaluation of traffic noise. In *NAG/DAGA 2009*, pages 984–987, 2009.

[43] M. Rychtarikova and G. Vermeir. Soundscape categorization on the basis of objective acoustical parameters. *Applied Acoustics*, 74:240–247, 2013.

[44] 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.

[45] A. Can and B. Gauvreau. Describing and classifying urban sound environments with a relevant set of physical indicators. *J. Ac. Soc. Am.*, 137:208, 2015.

[46] G. Rey Gozalo, J. Trujillo Carmona, J.M. Barrigon Morillas, R. Vilchez-Gomez, and V. Gomez Escobar. Relationship between objective acoustic indices and subjective assessments for the quality of soundscapes. *Applied Acoustics*, 97:1–10, 2015.

[47] L. Yu and J. Kang. Modeling subjective evaluation of soundscape quality in urban open spaces: An artificial neural network approach. *J. Ac. Soc. Am.*, 126:1163–1174, 2009.

[48] 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.

[49] 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.

[50] J.-R. Gloaguen, A. Can, M. Lagrange, and J.-F. Petiot. Estimation du niveau sonore du trafic routier au sein de mixtures sonores urbaines par la factorisation en matrices non-négatives. In *14ème Congrès Français d'Acoustique (CFA)*, 2018.

[51] D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[52] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[53] C. Lavandier, P. Aumond, S. Gomez, and C. Domingues. Urban soundscape maps modelled with geo-referenced data. *Noise Mapping*, 3:278–294, 2016.

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

[55] D. Rethage, J. Pons, and X. Serra. A Wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[56] 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.

[57] M. Rossignol, G. Lafay, M. Lagrange, and N. Misdariis. SimScene : a web-based acoustic scenes simulator. In *1st Web Audio Conference (WAC)*, 2015.

[58] 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.

[59] G. Lafay. *Simulation de scènes sonores environnementales - Application à l'analyse sensorielle et l'analyse automatique*. PhD thesis, Ecole Centrale de Nantes, 2016.

[60] 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.

[61] 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.

[62] F. Gontier, C. Lavandier, P. Aumond, M. Lagrange, and J.-F. Petiot. Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques. *Acta Acust. unit. Acust.*, 2019.

[63] R. Paulsen. On the influence of the stimulus duration on psychophysical judgement of environmental noises taken in the laboratory. In *INTER-NOISE 1997*, 1997.

[64] G. Brambilla and L. Maffei. Responses to noise in urban parks and in rural quiet areas. *Acta Acust. unit. Acust.*, 92:881–6, 2006.

[65] S. Kuwano, J. Kaku, T. Kato, and S. Namba. The experiment on loudness in field and laboratory: An examination of the applicability of laeq to mixed sound sources. In *INTER-NOISE 1997*, 1997.

[66] B. De Coensel, D. Botteldooren, B. Berglund, M. Nilsson, T. De Muer, and P. Lercher. Experimental investigation of noise annoyance caused by high-speed trains. *Acta Acust. unit. Acust.*, 93:589–601, 2007.

[67] S. Namba and S. Kuwano. Measurement of habituation to noise using the method of continuous judgment by category. *J. Sound Vib.*, 127:507–11, 1988.

[68] 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.

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

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

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

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

[73] G. Lafay, E. Benetos, and M. Lagrange. Sound event detection in synthetic audio: analysis of the DCASE2016 task results. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[74] ISO 226:2003. Acoustics - normal equal-loudness-level contours. Standard, International Organization for Standardization, Geneva, CH, 2003.

[75] J. Fritz, M. Elhilali, S. David, and S. Shamma. Auditory attention - focusing the searchlight on sound. *Curr. Opin. Neurobiol.*, 17:1–19, 2007.

[76] E. Kaya and M. Elhilali. Modelling auditory attention. *Phil. Trans. R. Soc.*, 372, 2016.

[77] C. Kayser, C. Petkov, M. Lippert, and N. Logothetis. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, 15:1943–1947, 2005.

[78] B. De Coensel and D. Botteldooren. A model of saliency-based auditory attention to environmental sound. In *20th International Congress on Acoustics (ICA)*, 2010.

[79] E. Kaya and M. Elhilali. A temporal saliency map for modeling auditory attention. In *46th Annual Conference on Information Sciences and Systems (CISS)*, 2012.

[80] N. Huang and M. Alhilali. Auditory salience using natural soundscapes. *J. Ac. Soc. Am.*, 141:2163–2176, 2017.

[81] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and models*. Springer, 1990.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

[87] 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.

[88] 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.

[89] 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.

[90] S. Kapka and M. Lewandowski. Sound source detection, localization and classification using consecutive ensembles of CRNN models. In *2019 Workshop on the Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019.

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

[92] K. Piczak. ESC: dataset for environmental sound classification. In *23rd ACM international conference on Multimedia (MM)*, pages 1015–1018, 2015.

[93] 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.

[94] A. Maas, A. Hannun, and A. Ng. Rectifier nonlinearities improve neural network acoustic models. In *30th International Conference on Machine Learning (ICML)*, 2013.

[95] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *A Field Guide to Dynamical Recurrent Networks*, 2001.

[96] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

[97] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: encoder-decoder approaches. In *8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, pages 103–111, 2014.

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

[99] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *2017 International Conference on Computer Vision (ICCV)*, 2017.

[100] J. Cramer, H.-H. Wu, J. Salamon, and J.P. Bello. Look, listen and learn more: design choices for deep audio embeddings. In *2019 IEEE International Conference on Audio, Signal and Speech Processing (ICASSP)*, 2019.

[101] A. Jansen, M. Plakal, R. Pandya, D. Ellis, S. Hershey, J. Liu, R. Moore, and R. Saurous. Unsupervised learning of semantic audio representations. In *2018 IEEE International Conference on Audio, Signal and Speech Processing (ICASSP)*, 2018.

[102] Y.-A. Chung and J. Glass. Learning word embeddings from speech. In *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

[103] T. Mikolov, I. Sutskever, K. Shen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *26th Conference on Neural Information Processing Systems (NIPS)*, 2013.

[104] Y.-A. Chung and J. Glass. Speech2Vec: A sequence-to-sequence framework for learning word embeddings from speech. In *Interspeech 2018*, 2018.

[105] M. Tagliasacchi, B. Gfeller, F. de Chaumont Quitry, and D. Roblek. Pre-training audio representations with self-supervision. *IEEE Signal Processing Letters*, 27:600–604, 2020.

[106] S. Pascual, M. Ravanelli, J. Serra, A. Bonafonte, and Y. Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. In *Interspeech 2019*, 2019.

[107] S. Ioffe and C. Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning (ICML)*, 2015.

[108] 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.

[109] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. Saurous, Y. Agiomvrgiannakis, and Y. Wu. Natural TTS synthesis by conditioning wavenet on Mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[110] J. Engel, L. Hantrakul, C. Gu, and A. Roberts. DDSP: differentiable digital signal processing. In *2020 International Conference on Learning Representations (ICLR)*, 2020.

[111] L. Boucheron, P. De Leon, and S. Sandoval. Low bit-rate speech coding through quantization of Mel-frequency cepstral coefficients. *IEEE Transactions on Audio, Speech and Language Processing*, 20:610–619, 2012.

[112] G. Min, X. Zhang, J. Yang, X. Zou, and Z. Pan. Speech reconstruction from MFCC based on nonnegative and sparse priors. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E98.A:1540–1543, 2015.

[113] G. Min, X. Zhang, J. Yang, and X. Zou. Speech reconstruction from Mel-frequency cepstral coefficients via l1-norm minimization. In *2015 IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2015.

[114] C. Donahue, J. McAuley, and M. Puckette. Adversarial audio synthesis. In *2019 International Conference on Learning Representations (ICLR)*, 2019.

[115] P. Neekhara, C. Donahue, M. Puckette, S. Dubnov, and J. McAuley. Expediting TTS synthesis with adversarial vocoding. In *Interspeech 2019*, pages 186–190, 2019.

[116] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis. Parallel WaveNet: fast high-fidelity speech synthesis. In *35th International Conference on Machine Learning (ICML)*, 2017.

[117] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio. SampleRNN: an unconditional end-to-end neural audio generation model. In *2017 International Conference on Learning Representations (ICLR)*, 2017.

[118] R. Prenger, R. Valle, and B. Catanzaro. Waveglow: a flow-based generative network for speech synthesis. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[119] P. Dhariwal, H. Jun, C. Payne, J.W. Kim, A. Radford, and I. Sutskever. Jukebox: a generative model for music. *ArXiv Preprints*, 2020.

[120] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *2017 Conference Neural Information Processing Systems (NIPS)*, 2017.

[121] B. De Coensel, D. Botteldooren, and T. De Muer. 1/f noise in rural and urban soundscapes. *Acta Acust. unit. Acust.*, 89:287–295, 2003.

[122] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama. Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency. In *International Conference on Digital Audio Effects (DAFx)*, pages 397–403, September 2010.

[123] D. Griffin and J. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Audio, Speech and Signal Processing (TASSP)*, 32:236–243, 1984.

[124] R. Huber and B. Kollmeier. Pemo-Q–a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, 14:1902–1911, 2006.

[125] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Trans-*

actions on Audio, Speech and Language Processing (TASLP), 19:2046–2057, 2011.

[126] E. Vincent. Improved perceptual metrics for the evaluation of audio source separation. In *10th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2012)*, 2012.

[127] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, 2016.

[128] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[129] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[130] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability and variation. In *2018 International Conference on Learning Representations (ICLR)*, 2018.

[131] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *2019 International Conference on Learning Representations (ICLR)*, 2019.

[132] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *27th Conference on Neural Information Processing Systems (NIPS 2014)*, 2014.

[133] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *34th International Conference on Machine Learning (ICML)*, pages 214–223, 2017.

[134] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. In *30th Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[135] P. Isola, J.-Y. Zhu, T. Zhou, and A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[136] P. Manocha, A. Finkelstein, R. Zhang, N. Bryan, G. Mysore, and Z. Jin. A differentiable perceptual audio metric learned from just noticeable differences. In *Interspeech 2020*, pages 2852–2856, 2020.

Titre : Analyse et synthèse de scènes sonores urbaines par approches d'apprentissage profond

Mots clés : Paysages sonores, Réseaux de capteurs acoustiques, Perception de sources sonores, Synthèse sonore

Résumé : L'avènement de l'Internet des Objets (IoT) a permis le développement de réseaux de capteurs acoustiques à grande échelle, dans le but d'évaluer en continu les environnements sonores urbains. Dans l'approche de paysages sonores, les attributs perceptifs de qualité sonore sont liés à l'activité de sources, quantités d'importance pour mieux estimer la perception humaine des environnements sonores. Utilisées avec succès dans l'analyse de scènes sonores, les approches d'apprentissage profond sont particulièrement adaptées pour prédire ces quantités. Cependant, les annotations nécessaires au processus d'entraînement de modèles profonds ne peuvent pas être directement obtenues, en partie à cause des limitations dans l'information enregistrée par les capteurs nécessaires pour assurer le respect de la vie privée.

Pour répondre à ce problème, une méthode pour l'annotation automatique de l'activité des sources d'intérêt sur des scènes sonores simulées est proposée. Sur des données simulées, les modèles d'apprentissage profond développés atteignent des performances « état de l'art » pour l'estimation d'attributs perceptifs liés aux sources, ainsi que de l'agrément sonore. Des techniques d'apprentissage par transfert semi-supervisé sont alors étudiées pour favoriser l'adaptabilité des modèles appris, en exploitant l'information contenue dans les grandes quantités de données enregistrées par les capteurs. Les évaluations sur des enregistrements réalisés in situ et annotés montrent qu'apprendre des représentations latentes des signaux audio compense en partie les défauts de validité écologique des scènes sonores simulées.

Dans une seconde partie, l'utilisation de méthodes d'apprentissage profond est considérée pour la resynthèse de signaux temporels à partir de mesures capteur, sous contrainte de respect de la vie privée. Deux approches convolutionnelles sont développées et évaluées par rapport à des méthodes état de l'art pour la synthèse de parole.

Title : Analysis and synthesis of urban sound scenes using deep learning techniques

Keywords : Soundscape, Acoustic sensor networks, Sound source perception, Sound synthesis

Abstract: The advent of the Internet of Things (IoT) has enabled the development of large-scale acoustic sensor networks to continuously monitor sound environments in urban areas. In the soundscape approach, perceptual quality attributes are associated with the activity of sound sources, quantities of importance to better account for the human perception of its acoustic environment. With recent success in acoustic scene analysis, deep learning approaches are uniquely suited to predict these quantities. Though, annotations necessary to the training process of supervised deep learning models are not easily obtainable, partly due to the fact that the information content of sensor measurements is limited by privacy constraints.

To address this issue, a method is proposed for the automatic annotation of perceived source activity in large datasets of simulated acoustic scenes. On simulated data, trained deep learning models achieve state-of-the-art performances in the estimation of source-specific perceptual attributes and sound pleasantness. Semi-supervised transfer learning techniques are further studied to improve the adaptability of trained models by exploiting knowledge from the large amounts of unlabelled sensor data. Evaluations on annotated in situ recordings show that learning latent audio representations of sensor measurements compensates for the limited ecological validity of simulated sound scenes.

In a second part, the use of deep learning methods for the synthesis of time domain signals from privacy-aware sensor measurements is investigated. Two spectral convolutional approaches are developed and evaluated against state-of-the-art methods designed for speech synthesis.