COLLEGE ED SCIENCES
DOCTORAL DE L'INGENIERIE
PAYS DE LA LOIRE ET DES SYSTEMES

CENTRALE NANTES

MEMOIRE DE DOCTORAT DE

L'ECOLE CENTRALE DE NANTES

ÉCOLE DOCTORALE N° 602
Sciences de l'Ingénierie et des Systèmes
Spécialité : *Signal, Image, Vision*

Par

Modan TAILLEUR

Large-Scale Monitoring of Acoustic Sources Based on Data-Driven Approaches for Urban Soundscape Analysis, Visualization, and Summarization

Projet de recherche doctoral présenté et soutenu à Nantes, le 20 Novembre 2025
Unité de recherche : UMR 6004, Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Rosa Maria Alsina-Pagès	Full professor, La Salle–Universitat Ramon Llull, Espagne
Nicolas Farrugia	Professeur de l'École des Mines–IMT, HDR, IMT Atlantique, Brest

Composition du Jury :

Président :	Romain Sérizel	Professeur des universités, Université de Lorraine, Nancy
Examinateurs :	Rosa Maria Alsina-Pagès	Full professor, La Salle-Universitat Ramon Llull, Espagne
	Nicolas Farrugia	Professeur de l'École des Mines-IMT, HDR, IMT Atlantique, Brest
	Christophe Hurter	Professeur des universités, Ecole Nationale de l'Aviation Civile, Toulouse
Directeur de recherches doctorales :	Mathieu Lagrange	Directeur de Recherche, CNRS, École Centrale de Nantes
Co-enc. de recherches doctorales :	Pierre Aumond	Chargé de Recherche, Université Gustave Eiffel, Bouguenais
Co-enc. de recherches doctorales :	Vincent Tourre	Maître de Conférence, École Centrale de Nantes

Acknowledgement

First of all, I would like to thank Mathieu Lagrange, Pierre Aumond, and Vincent Tourre for placing their trust in me and for making this PhD possible. I am also deeply grateful for their involvement, availability, and support throughout these three years. Each of them brought a unique perspective, and together they made this PhD a truly enriching experience.

I would like to thank all the participants who took part in the various perceptual and expert evaluations carried out during this project. Their contributions played an important role in supporting and advancing our research.

I also thank the members of the Signal, Image, and Sound team at LS2N for their warm welcome and the enjoyable moments we shared. I am especially grateful to Vincent Lostanlen, Barbara Pascal, and Han Han for our discussions and their valuable insights throughout this thesis.

Many thanks to Ophélie Moreau for her help with UX and visual design.

Finally, I thank my family and loved ones for their support over these three years.

Table of Contents

1 Introduction **19**

1.1 General context 19

1.2 Research questions 22

1.3 Manuscript structure 25

2 Sound Source Classification for Soundscape Analysis Using Fast Third-Octave Bands Data from an Urban Acoustic Sensor Network **27**

2.1 Introduction 30

2.2 Transcoder 33

2.2.1 General description 33

2.2.2 Transcoder architecture 34

2.2.3 Learning approach 35

2.2.4 Baselines 36

2.2.5 Validation 37

2.3 Quantative performance analysis 38

2.3.1 Materials and Method 38

2.3.2 Results 39

2.4 Qualitative performance analysis 39

2.4.1 Methods 40

2.4.2 Results 41

2.5 Discussion 45

2.5.1 Advantages of using FTO measurements 45

2.5.2 Motivations for using perceived time of presence instead of duration or dominance 47

2.5.3 Limitations of acoustic indices compared to AI-based source predictions 48

2.5.4 Potential applications and limitations of PANN-FTO predictions . . 48

2.6 Conclusion of the chapter 49

3 SoundSourceViz: A Visual Analytics Tool for Soundscape-Oriented Urban Planning with Acoustic Sensor Networks 51
3.1 Introduction . . . 54
3.2 Related Work . . . 56
3.2.1 Urban visual analytics . . . 56
3.2.2 Soundscape Visualization . . . 57
3.3 Design Rationale . . . 60
3.3.1 Design constraints . . . 61
3.3.2 Task Analysis . . . 61
3.3.3 Design Goals . . . 62
3.4 Components . . . 63
3.4.1 Color scheme used in components for traffic, voices, and birds . . . 63
3.4.2 Glyph design for spatial exploration . . . 65
3.4.3 Data vases for geo-referenced temporal exploration . . . 67
3.4.4 2D data vases of data vases for precise temporal exploration . . . 68
3.4.5 Support for pre- and post-SSD comparisons: color hues and component adaptations . . . 69
3.5 Analytical workflow . . . 70
3.5.1 Step A: Geo-referenced, glyph-based overview of the soundscape. . . . 70
3.5.2 Step B: Geo-referenced temporal patterns visualization for the analysis of selected locations. . . . 72
3.5.3 Step C: Precise temporal patterns visualization for the analysis of selected locations. . . . 73
3.6 Evaluation . . . 73
3.6.1 Implementation details . . . 73
3.6.2 Lorient CENSE dataset processing . . . 74
3.6.3 Experts panel . . . 74
3.6.4 Guided Analysis . . . 75
3.6.5 Evaluation protocol . . . 75
3.6.6 Results . . . 77
3.7 Discussion . . . 79
3.7.1 Improving the spatial exploration . . . 79
3.7.2 Improving temporal analysis . . . 79
3.7.3 Framework generalization . . . 80

3.7.4 Extension of the Framework . . . 80
3.8 Conclusion of the chapter . . . 80

4 Audio Skim Generation for Environmental Audio Recordings 83
4.1 Introduction . . . 86
4.2 Methods . . . 89
4.2.1 General framework . . . 89
4.2.2 Main Events Identification (MEI) . . . 90
4.2.3 Sequencing of Main Events (SME) . . . 91
4.3 Experimental protocol . . . 95
4.3.1 Evaluation Metrics . . . 95
4.3.2 Dataset . . . 100
4.3.3 Reference Audio Skims . . . 101
4.3.4 Baselines . . . 101
4.3.5 Experiment settings . . . 101
4.3.6 Computational efficiency of the pipeline . . . 104
4.3.7 Evaluation protocol . . . 104
4.4 Results . . . 105
4.4.1 Objective evaluation . . . 105
4.4.2 Subjective evaluation . . . 107
4.5 Discussions . . . 108
4.5.1 Regarding the relevance of the objective metrics . . . 108
4.5.2 Privacy considerations . . . 109
4.5.3 Considering more complex algorithms for addressing the SME problem 110
4.5.4 Interactive refinement of the audio skim . . . 110
4.5.5 Audio skims from discontinuous recordings . . . 111
4.5.6 Cyclic audio skims . . . 111
4.5.7 Audio skims for temporal profiles . . . 111
4.5.8 Helping users interpret summaries . . . 112
4.6 Conclusion . . . 112

5 Discussion 115
5.1 PANN-FTO: new applications and perspectives for improvement . . . 115
5.1.1 Using FTOs and sound sources predictions for other applications than urban environmental sound recordings . . . 115

5.1.2 Revisiting classifier choice for transcoder-based sound source inference 116
5.2 Reconstructing audio waveforms from FTOs . . . 117
5.2.1 Diffusion-based spectrogram super-resolution for waveform reconstruction from FTOs . . . 118
5.2.2 Audio skim generation from FTOs . . . 120
5.2.3 Robustness of FTO speech privacy enforcement against spectrogram super-resolution algorithms . . . 120
5.3 Towards standardization of sound source presence prediction from acoustic measurements . . . 122
5.4 Rethinking soundscape assessment: the role of non-acoustic factors . . . 123
5.5 Public communication on soundscape . . . 124

6 Conclusion **127**

7 Contributions **129**
7.1 Scientific publications . . . 129
7.1.1 Publications and submissions in international journals . . . 129
7.1.2 Publications and submissions in international conferences . . . 131
7.1.3 Publications in international workshops . . . 134
7.2 Open-source libraries . . . 135
7.3 Dataset . . . 136

A Aggregation algorithm for the display of the polar area diagrams network **137**

Bibliography **138**

List of Figures

1.1 Lorient CENSE sensor grid. Triangles indicate sensors that recorded data during the three-year deployment period. 21
1.2 Overview of research questions investigated during the PhD project. 23

2.1 PINV transcoder architecture, to recover a 1s sample fine spectrogram from a 1s sample coarse spectrogram 34
2.2 CNN transcoder architecture, to recover a 1s sample PANN fine spectrogram from a 1s sample coarse spectrogram 34
2.3 PANN CNN transcoder trained with a teacher-student approach, using the Binary Cross-Entropy (BCE) loss function 35
2.4 Classifier trained with a teacher-student approach, to match PANN outputs 36
2.5 Comparison between a coarse spectrogram, a fine spectrogram, and the fine spectrogram generated by the PINV transcoder and the CNN transcoder for a 1-s audio file extracted from TAU Urban Acoustic Scenes 2020 Mobile dataset. 37
2.6 Pearson correlations between annotations and predictions for two datasets. 40
2.7 Clock graphs representing the estimated perceived time of presence of traffic, voices and birds, on a sensor close to a traffic street, a pedestrian street and a residential area. The source time of presence prediction is averaged per hour. The time period considered ranges from January 1st to March 1st of 2020, with a random data sampling rate of 14% per day across 67 sensors, resulting in a total of about 500 hours of recording. 42
2.8 Clock graphs representing the presence of music during a music festival, and on a regular summer day. The source presence likelihood is averaged per hour. The time period considered for Figure a) encompasses every sunday of July 2021, taking into account all data from 5 sensors, resulting in a total of about 200 hours of recording. The time period considered for Figure b) is the 8th of August 2021, taking into account all data from 5 sensors, resulting in a total of about 100 hours of recording. 43

2.9 Clock graphs representing the presence of church bells on January 2020 when the bells were operational, and on October 2020, when they were not. The source presence likelihood is averaged per minute. The time period considered for Figure a) is October 2020, taking into account all data from 1 sensor, resulting in a total of about 300 hours of recording. The time period considered for Figure b) is January 2020, taking into account all data from 1 sensor, resulting in a total of about 600 hours of recording. . . 43
2.10 Maps representing the time of presence of traffic, voices and birds over different time intervals. Each node has 3 colors, the area of each color being related to the time of presence of their respective sound source. The circles are not superposed, but should be seen as one circle with the area of each color within it representing the time of presence of each sound source. The time period considered ranges from January 1st to March 1st of 2020, with a random data sampling rate of 0.7% per day across 67 sensors, resulting in a total of about 600 hours of recording. 45
2.11 Map illustrating the predicted presence of civil defense sirens on the first Wednesday of each month at 11h45. The map displays the locations of sirens and their estimated coverage. The time period considered encompasses every first wednesday of the month at 11h45 between January and July of 2020, taking into account all data from 67 sensors, resulting in a total of about 5 hours of recording. 46
2.12 Differences between time of presence annotation procedure, and mean of detection annotation on a synthetic example. On this example, close traffic noise would have a perceived time of presence close to 1/1, and distant traffic noise a perceived time of presence close to 0/1, even if their mean of detection annotation would be extremely close. 47

3.1 Steps of our soundscape visual analytics framework: (A) The polar area diagram network enables spatial comparison and exploration; (B) The geo-referenced 3D data vases link the spatial and temporal dimensions of the acoustic sensors; (C) The 2D data vases ease the temporal comparisons. The system is shown in comparison mode for steps B and C: the left side of the data vases corresponds to the period before the soundscape design intervention; the right side represents the period after. 54
3.2 Lorient Noise Map from the 2013 NAP (Ville de Lorient 2013). 55

3.3 Related work in soundscape visualization (1/2). 58
3.3 Related work in soundscape visualization (2/2). 59
3.4 Colors related to the presence of each sound source, along with the final color chosen for our study.. Colors that received less than 2% of answers are not represented here, and are merged into the hashed area. 63
3.5 Polar area diagrams showing traffic (red), voices (yellow), and birds (green), across the 3 different period visualization modes. 66
3.6 Polar area diagrams and data vases, with 1, 2, or 3 sources displayed. . . . 66
3.7 Data vases showing traffic (red), voices (yellow), and birds (green) presence for two temporal patterns, across the different period visualization modes. 68
3.8 Analytical workflow of SoundSourceViz. The steps of the workflow, A, B, C, correspond to those presented in Figure 3.1. 70
3.9 Overview of the visual analytics framework as used during the expert analysis. For clarity, button sizes have been enlarged. Red circles indicate the identifiers of interaction buttons and are included solely for reference in the paper. 71
3.10 Questions for the guided analysis. Questions were asked in French but are translated here into English. Each figure illustrates a different task involving interpretation of sound presence across time, space, and source. 76

4.1 Pipeline for Automatic Audio Skim generation. $e_{t,p}$ represents the pth embedding for period t, selected by a Kmeans algorithm for which we take the medoid of each cluster. δ represents the indices of the audio segments of the sequence. 90
4.2 $Beta_z$ weight for different z values, with $a_\beta = 100$, and $k_\beta = 7$. 93
4.3 Illustration of the performance metric in the faithfulness/highlightness plane with different z values. The area under the blue curve represents the performance of the audio skim generation algorithm. 99
4.4 Objective faithfulness and highlightness for audio skim generation methods. Each point represents an average over the evaluation datasets, and error bars indicate 95% confidence intervals. Results are averaged over the 40 12h recordings in the evaluation set. For the *Random* and *Downsample* methods, results are also averaged over 20 different seeds. For *SkimEAR* method, each point corresponds to a different value of the z parameter. . . 106

4.5 Subjective faithfulness and highlightness for all audio skim generation methods. Each point represents an average over the evaluation datasets, and error bars indicate 95% confidence intervals. Results are averaged over the 4 24-hour recordings in the subjective evaluation set. For the *Expert reference* and *SkimEAR* methods, each point corresponds to a different value of the z parameter. 108

5.1 Pipeline for audio super-resolution. Module outlined with dashes does not require any training, module with a plain outline is only pre-trained, and module outlined in bold is specifically trained for the task. 118
5.2 Pipeline for sound analysis, using both PANN-FTO and Diffspec. 119

A.1 Polar area diagram network without and with glyph aggregation. In (b), the number of sensors hidden beneath a glyph is specified within a black circle. 138

List of Tables

2.1 Differences between PANNs (ResNet38) and Cense spectral representations 31
2.2 Specifications of the different inference methods. The need for training annotations refers to the need to use time of presence annotations for each sound source 32

4.1 Overview of recordings with environment types, locations, dates, and dataset split. 100
4.2 MEI and SME parameter description, tuning outcomes, and selected values 103
4.3 Comparison of audio skim generation methods for the 40 12-hour recordings of the evaluation set, with 95% confidence interval. Bold values indicate results not significantly different from the best-performing condition, according to paired t-test ($p < 0.05$). 106
4.4 Comparison of the objective and subjective performances of audio skim generation methods for the 4 24-hour recordings of the subjective evaluation set, with 95% confidence interval. Bold values indicate results not significantly different from the best-performing condition, according to paired t-test ($p < 0.05$). 107

5.1 Mapping of sound event labels from the neonatal intensive care unit (NICU) to AudioSet. 116

Acronyms

AL Attack Level.

ASR Automatic Speech Recognition.

FTO Fast Third-Octave.

HSR Human Speech Recognition.

MEI Main Events Identification.

NICU Neonatal Intensive Care Unit.

PANNs Pretrained Audio Neural Networks.

PINV Pseudo-INVerse.

SD Source Diversity.

SME Sequencing of Main Events.

SSD SoundScape Design.

TC Temporal Consistency.

WER Word Error Rate.

Glossary

Attack Level Level of privacy attack regarding speech content. AL0 is an unintentional general audio reconstruction attack, AL1 targets generic speech recovery from unknown speakers, and AL2 focuses on reconstructing a specific individual's voice using their available speech data..

Audio Skim A short audio summary generated from a 24h audio recording, consisting of concatenated segments of the original audio.

AudioSet A large-scale dataset of over 2 million 10s audio clips drawn from YouTube videos, annotated with a hierarchical ontology of sound events.

BEATs Pre-trained audio classifier with acoustic tokenizers (S. Chen et al. 2022), trained on Audioset.

CNN-TrainSynth Source presence prediction model developped in (Gontier, Lavandier, et al. 2019), which predicts the time of presence of traffic, voices, and birds using a convolutional neural network.

Data Vase A 3D visualization technique developed in (Thakur et al. 2010) for exploring geo-referenced temporal data. Each "vase" encodes a value of interest in its radius, with the vertical axis representing time.

Diffspec A diffusion-based spectrogram super-resolution algorithm, described in (Tailleur, Benaatia, et al. 2025).

Fast Third-Octave A compressed spectral representation with 20–29 third-octave band levels recorded every 125ms.

FTO-TO-WAV A deep learning super-resolution pipeline that transforms Fast Third-Octave (FTO) data into audio waveforms, using the Diffspec super-resolution architecture trained on TAU Urban Acoustic Scenes 2020 (Mesaros, Heittola, Diment, et al. 2017) and described in (Tailleur, Benaatia, et al. 2025).

GRAFIC Audio dataset created from soundwalks in Paris's 13th district, with perceptual annotations of pleasantness, liveliness, loudness, and time of presence (Aumond, Can, De Coensel, et al. 2017).

L-CLAP embeddings Embeddings obtained from the CLAP (Contrastive Language–Audio Pretraining) model trained by LAION (Wu et al. 2023).

Lorient Cense A dense urban sensor network deployed in Lorient, France, which recorded FTOs on 78 sensors for over 3 years.

Lorient-1k Dataset of 30 acoustic scenes (45s each) recorded at 10 locations in Lorient, France, totaling 22.5min, with expert annotations of perceived time of presence for birds, traffic, and voices (Gontier, Lostanlen, et al. 2021).

PANN-FTO The PANN ResNet38 model applied to Fast Third-Octave (FTO) data transcoded into the Mel spectrogram format.

PANN-Mel The PANN ResNet38 model applied directly to Mel spectrograms with 10-ms hops generated from audio waveforms.

Pretrained Audio Neural Networks Deep learning models designed for general-purpose audio tagging and sound event classification.

Pseudo-Inverse Method A signal processing technique to estimate missing or under-sampled spectral components using the Moore–Penrose pseudo-inverse (Penrose 1955).

SONYC-UST Dataset of urban sound recordings from the Sounds of New York City (SONYC) project (Cartwright et al. 2019), containing 3,068 recordings with fine-grained urban sound labels.

SoundScape Design Urban planning approaches aimed at shaping acoustic environments.

TAU Urban Acoustic Scenes 2020 A publicly available dataset for acoustic scene classification (Mesaros, Heittola, and Virtanen 2018).

Time of Presence A perceptual metric describing the overall presence of a sound source in a soundscape, used as an alternative to "dominance," reflecting holistic human judgment rather than physical onset/offset annotations.

Transcoder Model converting FTOs into Mel spectrograms compatible with pre-trained audio classifiers such as PANNs.

UrbanSound8K Dataset of 8,732 labeled urban sound excerpts (around 4s each) from 10 classes including car horn, drilling, and dog bark (Salamon et al. 2014).

Chapter 1

Introduction

1.1 General context

The sound environment plays a significant role in the quality of life and can impact both physical and mental health. The European Environment Agency, in their newest report (European Environment Agency 2025), reveals that at least one in five Europeans lives in areas where noise exposure exceeds healthy limits. Traffic noise pollution, for instance, has been consistently associated with cardiovascular diseases (Hoffmann et al. 2006; Stansfeld et al. 2005) and elevated stress levels (Babisch et al. 2001). Conversely, exposure to natural sounds has been shown to positively affect well-being (Buxton et al. 2021; Zhu et al. 2024).

With rising awareness of the adverse health effects of environmental noise, the European Union adopted the Directive 2002/49/EC (European Commission 2002). This directive establishes a common European framework for assessing and managing environmental noise, particularly from road, rail, air traffic, and industrial sources. Its main objectives are to monitor noise exposure, inform the public, and guide cities in developing action plans to prevent or reduce harmful effects. In particular, it requires European cities of more than 100,000 inhabitants to maintain publicly available noise maps. These noise maps are generated using acoustic modeling techniques that estimate noise levels based on simulated noise emissions and their propagation. The models take into account various factors such as weather conditions, topography, and ground absorption. While modeled noise maps offer an initial overview of the sound environment, they often fail to faithfully reflect actual on-site acoustic conditions. These maps typically describe average situations, which do not capture the event-based nature of certain sound sources and the associated disturbances. As a result, sudden and irregular events such as honking, emergency vehicle sirens, loud motorcycles, delivery activities, or repeated noise peaks from aircraft or rail traffic, are difficult to model accurately. Modeling is also particularly challenging in dense urban areas where traffic speeds are poorly defined due to frequent

acceleration and braking caused by traffic lights or congestion. In such contexts, in-situ noise measurements are commonly conducted, as they provide more reliable and realistic insights into the nuisances actually experienced by residents.

Depending on the noise measurement objectives, acoustic sensors may capture raw noise levels, record audio directly, or collect spectro-temporal representations of the audio. A common form of such representations is the third-octave bands format, which provides sound levels across 20 to 29 frequency bands at regular time intervals. Fast Third-Octave (FTO) representations, in particular, are typically computed every 125ms. The IEC 61672-1 (IEC 2013) standardizes the measurement of these representations, which find application in various noise monitoring contexts (Aumond, Can, De Coensel, et al. 2017; Can, Picaut, et al. 2021; Farrés 2015; Mietlicki et al. 2015; Nilsson et al. 2007; Torija et al. 2013). FTOs present several advantages in long-term monitoring. Most notably, they enforce unintelligibility, thus preserving privacy, as demonstrated in (Gontier, Lagrange, et al. 2017). Furthermore, these representations are lightweight, with a bit rate 140 times lower than that of mono waveform recordings (16-bit - 32 kHz). They are also widely accessible, as most sound level meters natively support their computation.

The recent technological advent of low cost microphones and innovative airborne data transfer protocols allows the deployment of acoustic sensor networks. Some large cities have deployed those networks in the last decade, in order to perform continuous monitoring of urban environments at different locations (Abeßer et al. 2019; Can, Picaut, et al. 2021; Dove et al. 2022; Farrés 2015; Mietlicki et al. 2015; Sevillano et al. 2016). These networks typically comprise dozens of sensors, and record data for durations ranging from several days to multiple years, which is stored on central servers. Sensors are usually installed on existing urban infrastructure, such as street lamps or balconies. FTO recordings are particularly well-suited to these networks, as they provide a richer spatio-temporal representation of the sound environment than simple sound level measurements, while remaining lightweight enough to allow the storage of large datasets with limited space requirements.

In particular, the Lorient Cense sensor network (Can, Picaut, et al. 2021), deployed in the city of Lorient, France, is one of the largest and the longest-running sensor network of its kind. The network consists of 78 noise sensors distributed across the urban area, with an average spacing of 50m, covering approximately 2.5km^2. Deployed between January 2020 and April 2022, the sensors are designed to record FTO data continuously. The entire system is connected to the cloud and leverages the city's public street lighting infrastruc-

ture for deployment, using power-line communication for data transmission. Each sensor is installed at a height of 4m above ground level. The sensors utilize micro-electromechanical system (MEMS) microphones, equipped with a Raspberry Pi for recording and transmitting purposes, featuring real-time audio processing capabilities. They allow the recording of FTOs using 29 third-octave bands covering the frequency range from 20 Hz to 12.5 kHz. The geographical distribution of the sensors is shown in Figure 1.1. The complete dataset comprises 1.5TB of FTO data, corresponding to over 500,000h of measurements collected by the network. Of the 78 sensors, 48 were operational at least 75% of the days between January and June 2020. Due to difficulties in maintaining the network, this number decreased to 24 sensors between June 2020 and January 2021, 8 sensors between January 2021 and June 2021, and 1 sensor between June 2021 and January 2022.

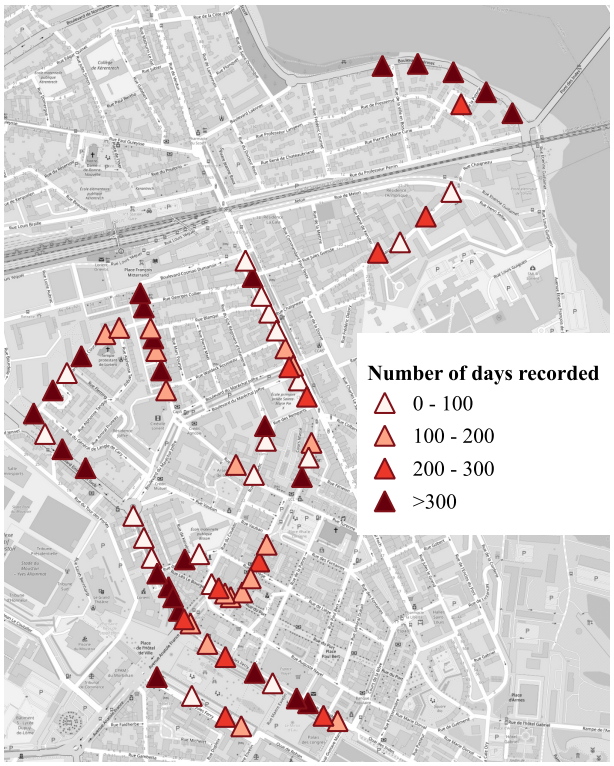


Figure 1.1 – Lorient CENSE sensor grid. Triangles indicate sensors that recorded data during the three-year deployment period.

The most common approach to analyze the sound environment, using either simulations used to produce noise maps or acoustic sensors, is to examine overall noise levels (dBA) and determine whether they exceed defined thresholds across different locations and time periods. For example, The World Health Organization suggests that healthy

noise levels should not surpass 55 decibels during the day and 40 at night (World Health Organization 2011). Noise levels thus provide useful information about environmental conditions, as they are strongly correlated with perceived annoyance (Nilsson et al. 2007). However, more recent approaches have shifted away from viewing sound solely as a nuisance to be mitigated. These approaches revolve around the concept of "soundscape", which is defined in the ISO/DIS 12913-2 (International Standard 2018) as the “acoustic environment as perceived or experienced and/or understood by a person or people, in context”. This conceptual shift reflects a broader interest in understanding how people experience their acoustic environment, considering not only unwanted or disturbing sounds, but also those that are pleasant, engaging, or restorative. From this perspective, sound is treated as a resource to be shaped and enhanced, rather than an unwanted byproduct of urban activity.

Soundscape quality is typically assessed through questionnaires. These surveys focus on perceptual attributes such as pleasantness, eventfulness, and annoyance, as well as the perceived presence or dominance of various sound sources (e.g., road traffic, construction work, socially or naturally connoted sources). However, such evaluations are highly localized in both time and space, and their implementation is labor-intensive and difficult to scale. As an alternative, recent work suggests that sensor-based data can be used to approximate soundscape quality in a more scalable and continuous manner (Aumond, Can, De Coensel, et al. 2017; Gontier, Lavandier, et al. 2019; Hou et al. 2023; Siliezar et al. 2024). Beyond monitoring, such acoustic data could also support new ways of interpreting and engaging with the soundscape.

In this manuscript, we explore new approaches to soundscape analysis, focusing on methods based on acoustic measurements. The next section introduces the three research questions addressed in this doctoral work.

1.2 Research questions

This manuscript addresses three research questions (RQ) that were explored throughout the PhD project. An overview of these questions and the corresponding contributions is presented in Figure 1.2.

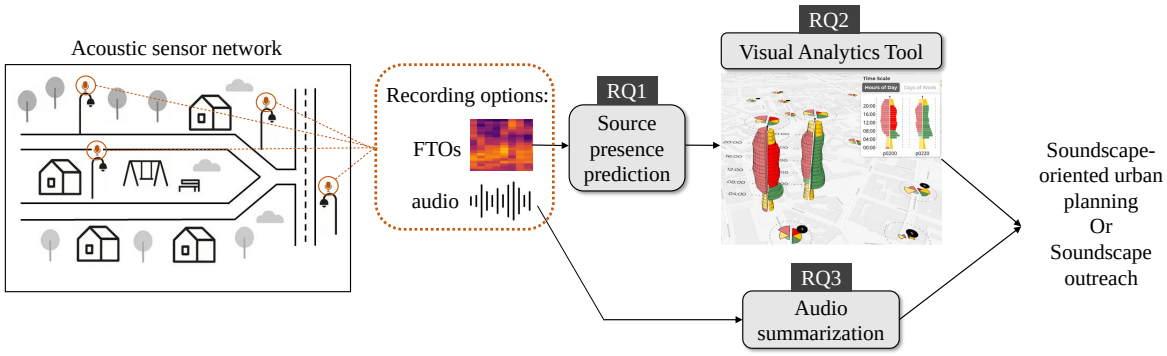


Figure 1.2 – Overview of research questions investigated during the PhD project.

RQ1

To what extent can the presence of sound sources be predicted from fast third-octaves recorded by acoustic sensors?

Usually, soundscape studies heavily rely on the concepts of annoyance and pleasantness to describe the quality of a sound. However those concepts are highly context dependant. Indeed, a sound environment very animated with lots of voices can for example be perceived as pleasant by tourists, but highly disturbing by local citizens. Other studies rather propose to estimate the perceived presence of various sound sources within an environment. In particular, road traffic, human voices, and bird sounds have been frequently considered in soundscape evaluations (Aumond, Can, De Coensel, et al. 2017; Ricciardi et al. 2015), representing the prevalent sounds in mechanical, human, and natural urban environments (International Standard 2014; Jeon et al. 2015). Previous work argues that focusing on characterizing these three environments is essential to better describe a soundscape (Aletta et al. 2016; Axelsson et al. 2010; Botteldooren, Lavandier, et al. 2011).

While the presence of sound sources has often been estimated through listening tests or questionnaires, such evaluations are resource-intensive and hardly scalable. This limitation has motivated the development of automatic approaches for estimating source presence, particularly from FTOs, whose advantages were discussed in the previous section. Gontier, Lavandier, et al. (2019) developed a deep learning model to predict the perceived presence of traffic, voices, and birds from FTOs. However, their model lacked robustness when applied to datasets that differed from the specific conditions (e.g., city context, sensor

height) used during training. In Chapter 2, we present a new method for predicting the time of presence of traffic, voices, and birds, and demonstrate that our predictions remain reliable across multiple datasets.

RQ2

How can a visual analytics tool support soundscape-informed urban planning using acoustic sensor data?

Urban planning approaches known as Soundscape Design (SSD) strategies (Bild et al. 2016) focus not only on mitigating unwanted noise, but also on promoting pleasant sounds through environmental modifications, or acoustic installations (Asdrubali et al. 2014; Hellström et al. 2014; Licitra et al. 2010). SSD strategies make use of diverse types of information, such as inhabitants, space users or stakeholders opinions, acoustic measurements, and soundscape simulations. They often involve presenting soundscape redesign proposals to interested parties, followed by the implementation of both active and passive interventions aimed at transforming the sonic environment (De Coensel et al. 2010).

Urban acoustic sensing offers the potential to enrich SSD approaches with continuous, large-scale data. Sensor networks provide fine-grained temporal and spatial information about the acoustic environment, enabling a more comprehensive understanding of urban soundscapes. However, the complexity and volume of this data pose challenges for urban planners and decision-makers, who may lack tools to interpret such information effectively. To address this gap, there is a need to develop visual analytics tools specifically tailored to soundscape-informed urban planning. These tools should enable users to intuitively explore, compare, and interpret acoustic data and perceptual predictions derived from sensor networks.

In Chapter 3, a soundscape-oriented visual analytics tool is presented. It is designed to assist urban planners in analyzing sensor network data and supporting soundscape-informed decision-making.

RQ3

How to provide meaningful audio skims of long audio recordings?

Acoustic measurements and even high-level perceptual attributes, such as pleasantness

or the presence of specific sound sources, are not always intuitive for understanding the quality of a sound scene. Providing users with audio playback is an effective way to convey these insights more directly. However, in networks such as the Lorient Cense system, each sensor records thousands of days of audio, making exhaustive listening impractical and overwhelming. Generating short, representative audio summaries for each location could offer a more accessible way to understand the soundscape of a location.

In the video domain, a video skim refers to a summary created by concatenating selected segments of a video (Apostolidis, Adamantidou, et al. 2021; Sundaram et al. 2002). In Chapter 4, we extend the concept of video skimming to the audio domain, while accounting for the specificities of environmental noise monitoring. We propose a method for generating short audio skims from long audio recordings. Although both input and output durations are flexible, a typical use case is summarizing 24h of audio into a 1min skim. The method also lets users adjust the balance between preserving the background sound and chronology of the original recording and emphasizing its most salient events.

1.3 Manuscript structure

Each research question is examined in Chapter 2, Chapter 3, and Chapter 4, respectively. These chapters are based on research articles that have been published (Chapter 2), submitted for peer review (Chapter 3), or that are in preparation (Chapter 4) for submission to international journals. These main contributions are supported by additional work, either published or submitted to international conferences or workshops, which will not be detailed in this manuscript. These supporting materials are considered as either prior knowledge or complementary contributions.

Over the course of the doctoral work, RQ1 was addressed first, while interviews were conducted in preparation for tackling RQ2. The investigations for RQ2 and RQ3 were then carried out in parallel. As a result of this chronology, the audio playback feature proposed in Chapter 4 is not included in the visual analytics tool presented in Chapter 3. Integrating such a feature could be a valuable direction for future work on RQ2.

Each main chapter includes its own introduction, state of the art, and discussion, which explains the absence of a comprehensive general introduction to this manuscript.

Chapter 2

Sound Source Classification for Soundscape Analysis Using Fast Third-Octave Bands Data from an Urban Acoustic Sensor Network

ABSTRACT

The exploration of the soundscape relies strongly on the characterization of the sound sources in the sound environment. Novel sound source classifiers called Pre-trained Audio Neural Networks (PANNs) (Kong et al. 2020), are capable of predicting the presence of over 500 diverse sound sources. Nevertheless, PANNs models use fine Mel spectro-temporal representations as input, whereas sensors of an urban noise monitoring network often record FTOs, which have significantly lower spectro-temporal resolution. In (Tailleur, Lagrange, et al. 2023), we developed a transcoder to transform FTOs into the fine Mel spectro-temporal representation used as input of PANNs. In this chapter, we demonstrate that employing PANNs with fast third-octaves data, processed through this transcoder, does not strongly degrade the classifier's performance in predicting the perceived time of presence of sound sources. Through a qualitative analysis of a large-scale fast third-octave dataset, we also illustrate the potential of this tool in opening new perspectives and applications for monitoring the soundscapes of cities.

Context

The content of this chapter is based on the main-authored scientific article: "Sound Source Classification for Soundscape Analysis Using Fast Third-Octave Bands Data from an Urban Acoustic Sensor Network" (Tailleur, Aumond, Lagrange, et al. 2024), published in the Journal of the Acoustical Society of America (JASA).

MAIN CONTRIBUTION

Reference: Modan Tailleur, Pierre Aumond, Mathieu Lagrange, and Vincent Tourre (2024), « Sound source classification for soundscape analysis using fast third-octave bands data from an urban acoustic sensor network », *in*: *The Journal of the Acoustical Society of America (JASA)*

Section 2.2 expands upon the original article by incorporating additional details from "Spectral Transcoder: Using Pretrained Urban Sound Classifiers on Undersampled Spectral Representations" (Tailleur, Lagrange, et al. 2023), published at the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE). For details on the evaluation procedure of this paper, readers are invited to consult the original publication.

SUPPORTING CONTRIBUTION

Reference: Modan Tailleur, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2023), « Spectral trancoder: using pretrained urban sound classifiers on undersampled spectral representations », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*

Gontier et al. (Gontier, Lavandier, et al. 2019) addressed multi-label classification tasks in urban environments using a Convolutional Neural Network (CNN) directly trained on third-octave spectrograms. While their model showed good performance on the Lorient Cense dataset (Can, Picaut, et al. 2021), it lacks robustness on other FTO datasets. This limitation arises partly from training the model on highly homogeneous datasets. Pre-trained models such as YAMNet and PANNs, on the other hand, have shown robustness in a variety of sound source classification tasks. Unfortunately, these models are trained on Mel spectrograms with a 10-ms frame step,

and can only consider the corresponding Mel representation as input.
This paper proposes a neural network that transcodes FTOs into Mel spectrograms, making them compatible with the input format of pre-trained classification models. This approach leads to a 62.4% accuracy on UrbanSound8k (Salamon et al. 2014) and 0.44 macro AUPRC on SONYC-UST (Cartwright et al. 2019). The transcoder is described in Chapter 2, Section 2.2, but its evaluation is not detailed exhaustively in this manuscript.

2.1 Introduction

Soundscape studies often rely on perceptual attributes such as annoyance or pleasantness. However, these attributes are highly context-dependent, making their assessment both difficult and potentially misleading without specifying to whom, in what situation, and for what reason a sound might be annoying or pleasant. An alternative approach focuses on estimating the perceived presence of traffic, voices, and birds, which represent mechanical, human, and natural environments. While such perceptual attributes are typically assessed through questionnaires, this work explores their estimation from FTOs.

Some acoustic indices, which can be computed from FTOs, have shown significant correlations with the presence of specific sound sources, such as LAeq for road traffic and the Time Frequency Second Derivation Index (TFSD) for bird presence (Aumond, Can, De Coensel, et al. 2017). The package seewave (Sueur et al. 2008) incorporates many other acoustic indices, designed for ecoacoustic analyses, which are intended to detect geophonic, anthropogenic, and natural sounds. Unfortunately, creating specific acoustic indices is time-consuming and case-specific as it requires a context-aware understanding of the acoustic characteristics of the target sound source. For example, while TFSD effectively captures rapid spectro-temporal variations in bird sounds within a particular frequency range, it does not perform well when applied to other sound sources with similar characteristics yet different frequency ranges, such as human voices (Aumond, Can, De Coensel, et al. 2017). Similarly, although LAeq correlates strongly with the time of presence of traffic noise in busy streets, its effectiveness in predicting the presence of road traffic diminishes in environments dominated by other sound sources, such as construction sites or factories.

Artificial intelligence has proven to be a valuable tool for detecting sound sources (Bansal et al. 2022) and predicting their perceived time of presence. Gontier at al. (Gontier, Lavandier, et al. 2019) obtained accurate predictions of the perceived time of presence of traffic, voices and birds, by training a Convolutional Neural Network (CNN) on FTOs. In this chapter, we refer to their model as **CNN-TrainSynth**. While CNN-TrainSynth shows good performance on the Lorient Cense dataset presented in Chapter 1, it also demonstrates a lack of robustness when applied to alternative datasets recorded with FTOs. This limitation stems, in part, from the model's training on highly homogeneous data. Furthermore, despite the standardization of time resolution for FTOs, variations in other parameters, such as the number of frequency bins and the frequency bounds, can

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

Table 2.1 – Differences between PANNs (ResNet38) and Cense spectral representations

induce problems in generalization to other data distribution. Additionally, while the architecture could easily be adapted to a given sound source, it would require re-training and thus a cumbersome procedure to produce a clean and annotated dataset. Other studies have tried to predict directly annoyance. Due to its strong context dependency, annoyance is particularly difficult to model in a transparent and interpretable way, making its direct prediction misaligned with the principles of explainable AI (Mueller et al. 2019). Hou et al. (Hou et al. 2023) addressed this challenge by jointly predicting annoyance and sound sources, illustrating an effort to improve interpretability of this high-level perceptual attribute.

Recent advances in deep learning have led to the development of robust and versatile pre-trained classifiers, known as Pre-trained Audio Neural Networks (PANNs) (Kong et al. 2020). PANNs are highly effective deep learning models that have been trained on Audioset, an extensive audio dataset comprising more than 2 million audio clips (Gemmeke et al. 2017). These models operate as sound classifiers and are capable of making predictions on the presence of 527 different sound sources without the need for additional training. For each class, they output a confidence score between 0 and 1, indicating the likelihood of the source's presence. Over 20 distinct pre-trained models are accessible online, all employing the same 10-ms 64Mel bands spectro-temporal representation as input. This representation has a much higher frequency and time resolution compared to FTOs, as shown in Table 2.1.

This chapter builds upon the work presented in (Tailleur, Lagrange, et al. 2023) which

	acoustic indicator	CNN-TrainSynth	PANN-FTO	PANN-Mel
input spectrogram	FTO	FTO	FTO	10-ms 64Mel
output classes	1	3	527	527
needs training annotations	Yes	Yes	No	No

Table 2.2 – Specifications of the different inference methods. The need for training annotations refers to the need to use time of presence annotations for each sound source

focused on introducing an algorithm for transforming FTOs into 10-ms 64Mel bands representations. This transcoding technique allows to provide FTO measurements to classification algorithms that necessitates Mel spectro-temporal representations as input, such as PANNs. In the remaining of this manuscript, the ResNet38 PANN model with 10-ms 64Mel bands spectrograms as input will be referred to as **PANN-Mel**, and the same PANN model that uses transcoded Mel spectrograms from FTO measurements will be referred to as **PANN-FTO**. A summary of the available models and methods to predict the time of presence of sound sources is presented in Table 2.2. Technical aspects of the transcoder are summarized in Section 2.2. Despite only being tested for this specific use case scenario, theoretically the transcoder can be used to convert any spectro-temporal representation into 10-ms 64Mel bands spectrograms. It was demonstrated that the transcoding technique applied to PANN on FTOs enabled satisfactory classification performances on two urban sound datasets: SONYC-UST (Cartwright et al. 2019) and UrbanSound8k (Salamon et al. 2014). However, the assertion that PANN-FTO output classes are highly correlated with perceptual evaluations of time of presence is yet unproven.

This study is specifically focused on showcasing the effectiveness of PANN in accurately predicting the perceived time of presence, a crucial metric in soundscape analysis, for diverse sound sources. It particularly assesses the applicability of PANN in analyzing the time of presence of datasets recorded with FTOs. PANN predicts accurately not only the time of presence for traffic, voices, and birds but also the presence of a diverse range of 527 different sound sources. Thanks to the transcoder developped in (Tailleur, Lagrange, et al. 2023), using PANN on FTOs allows predictions on lightweight datasets, while ensuring privacy. Consequently, this approach facilitates analysis on large datasets, such as Lorient Cense (explored in section 2.5) which contains more than 500,000h of FTO data on 75

different sensors.

In the forthcoming Section 2.2, we present the transcoding algorithm used for PANN-FTO. In Section 2.3, we elaborate on our findings concerning perceived time of presence assessment performance. We then show an analysis of the soundscape of Lorient through a FTO recorded dataset in Section 2.4. In section 2.5, we discuss the relevance, the opportunities, and potential enhancements arising from this approach. [1]

2.2 Transcoder

PANNs models (Kong et al. 2020) take a 10-ms 64Mel frequency band spectrogram as input, as shown in table 2.1. Audio can easily be transformed into the corresponding Mel spectrogram in order to use PANNs. Nonetheless, there is no easy transformation from FTOs to 10-ms 64Mel frequency band spectrograms, as the latter have much finer resolutions. In order to use PANN models to predict the presence of sound source from FTOs, a transcoder is employed to convert them into 10-ms 64Mel frequency bands spectral representations. In the subsequent sections, we will refer to the FTOs spectrograms as **"coarse spectrograms"** and the 10-ms 64Mel frequency bands spectrograms as **"fine spectrograms"**. We direct readers to (Tailleur, Lagrange, et al. 2023) for further details on the training procedure, and on the model's classification performances compared to other state-of-the-art methods.

2.2.1 General description

The transcoder outputs the fine spectrogram format that is required as input of PANNs models. This transformation is performed on audio segments with a duration of 1 second. The Lorient Cense project's method for FTO (coarse spectrogram) calculation (Can, Picaut, et al. 2021) is chosen. This method involves computing 29 third-octave bands within the frequency range of 20Hz to 12.5kHz, using a tukey 125-ms temporal window, as depicted in Table 2.1.

The transcoder is trained on the TAU Urban Acoustic Scenes 2020 Mobile dataset (Mesaros, Heittola, and Virtanen 2018), an urban dataset of raw audio data. This dataset does not include any annotations regarding the presence of sound sources in the audio recordings. It consists of 30h of 10-second audio clips from 10 different acoustic scenes.

1. Open source code is available at https://github.com/modantailleur/paperSoundscapeSourceClassification.

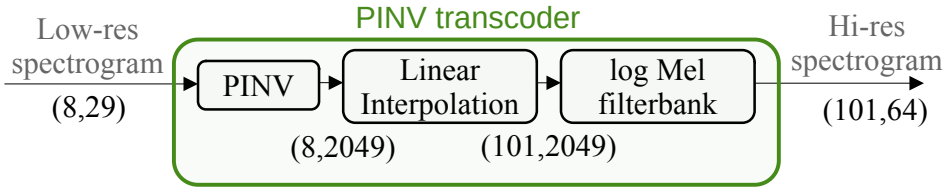


Figure 2.1 – PINV transcoder architecture, to recover a 1s sample fine spectrogram from a 1s sample coarse spectrogram

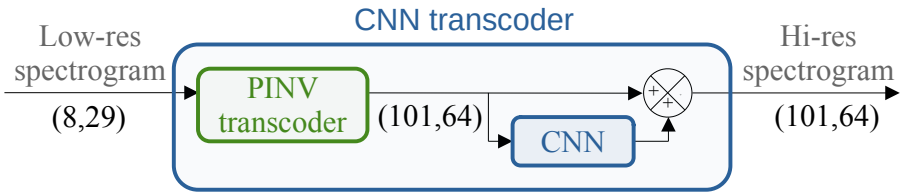


Figure 2.2 – CNN transcoder architecture, to recover a 1s sample PANN fine spectrogram from a 1s sample coarse spectrogram

2.2.2 Transcoder architecture

The proposed CNN transcoder model consists of two parts: a Pseudo-INVerse (PINV) transcoder and a Convolutional Neural Network (CNN), as shown in Figures 2.1 and 2.2. The PINV transcoder presented in Figure 2.1 first reconstructs the full-band spectrogram from the coarse spectrogram using a pseudo-inverse method (Penrose 1955). The PINV stage enables leveraging knowledge from the third-octave transform to provide an initial estimation of the fine spectrogram. This stage also theoretically allows the model to adapt to any coarse spectrogram calculated using different numbers of frequency bins or varying time weightings. The CNN stage improves the quality of this estimation by adding residual information to it, as illustrated in Figure 2.2. The model is very light, representing 0.3% of our selected ResNet38 PANN's total number of parameters.

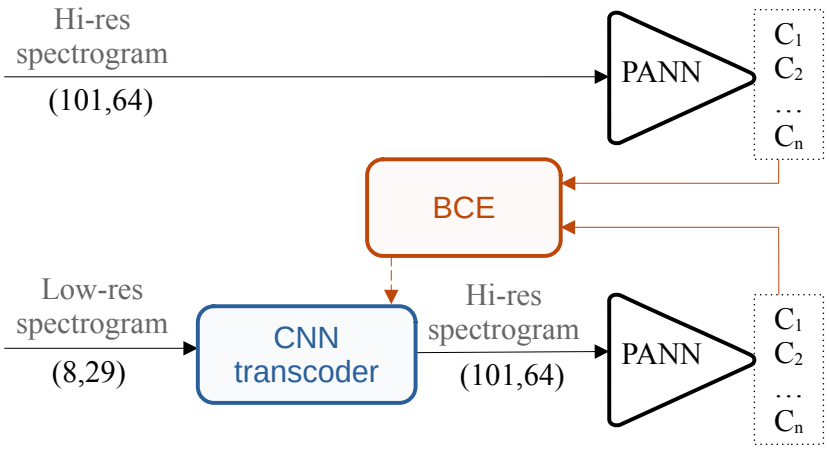


Figure 2.3 – PANN CNN transcoder trained with a teacher-student approach, using the Binary Cross-Entropy (BCE) loss function

2.2.3 Learning approach

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

The teacher-student approach taken to train the transcoder diverges slightly from the aforementioned regular teacher-student methods, while benefiting from the same advantages. Notably, only the parameters of the CNN transcoder are updated during training (see Figure 2.3), whereas the entire student model would be trained if a regular teacher-student approach had been employed. As a result, at the end of the transcoder training, the PANN model used in the student process remains identical to the one employed by the teacher. Training of the transcoder is performed using the Binary Cross-Entropy (BCE) loss function, computed on the predictions of both the teacher and student PANN classifiers.

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

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

2.2.4 Baselines

We compare the classification performance of the CNN-logits transcoder with the performance of a reference PINV transcoder (as shown in Figure 2.1), which does not require any learning. We also retrain a PANN model, as well as efficient nets (efficient net b_0 with 4,682,059 parameters and efficient net b_7 with 65,135,455 parameters) (Tan et al. 2021), using pseudo-inverted Mel spectrograms as input with the method illustrated in Figure 2.4. Those model serve as baselines for validating our PANN-FTO model.

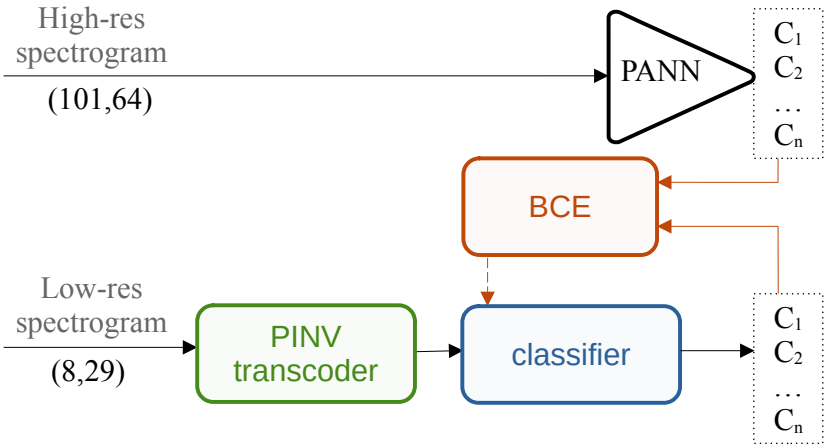


Figure 2.4 – Classifier trained with a teacher-student approach, to match PANN outputs

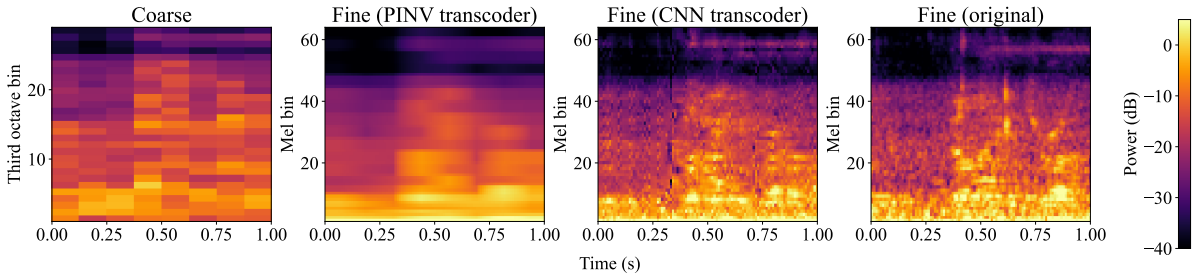


Figure 2.5 – Comparison between a coarse spectrogram, a fine spectrogram, and the fine spectrogram generated by the PINV transcoder and the CNN transcoder for a 1-s audio file extracted from TAU Urban Acoustic Scenes 2020 Mobile dataset.

2.2.5 Validation

The training of the model lasts approximately 4h on a V100 GPU. Interestingly, the resulting PANN-FTO model trained with a transcoder, outperforms the PANN and EfficientNet models trained with a regular teacher-student method. This surprising outcome is particularly noteworthy considering that the CNN transcoder is substantially smaller in size. PANN-FTO obtains an accuracy of 89.3% in predicting the same first class as the teacher PANN-Mel model on the evaluation dataset of TAU Urban Acoustic Scenes 2020 Mobile dataset (Mesaros, Heittola, and Virtanen 2018). Models trained using regular teacher-student methods only reached a maximum accuracy of 83.7%. The proposed model has also been tested in multi-class classification tasks and multi-label classification tasks, reaching a 62.4% accuracy on the UrbanSound8k dataset (Salamon et al. 2014), and a .44 mAUPRC on Sonyc-UST dataset (Cartwright et al. 2019), see (Tailleur, Lagrange, et al. 2023) for more details.

Figure 2.5 illustrates an instance of fine spectrogram generation from coarse representation, on a 1-second clip from the evaluation dataset. Although the transcoder is not explicitly trained to replicate a coherent fine spectrogram but rather to produce the same output classes as the target spectrogram, it yields a reconstructed spectrogram with notable fidelity to the original one.

These results concern multi-class and multi-label classification tasks, which differ somewhat from the estimation of perceived time of presence. In the upcoming sections, we will evaluate the algorithm's efficacy in predicting annotated times of presence across different datasets.

2.3 Quantative performance analysis

2.3.1 Materials and Method

Datasets

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

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

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

PANN prediction processing to predict the perceived time of presence

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

To represent the time of presence of traffic, voices, and birds, we select three output

classes from PANN that closely align semantically with these sources, namely "traffic noise, roadway noise", "speech", and "bird vocalization". For further details on the available PANNs output classes, please refer to the Audioset ontology (Gemmeke et al. 2017).

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

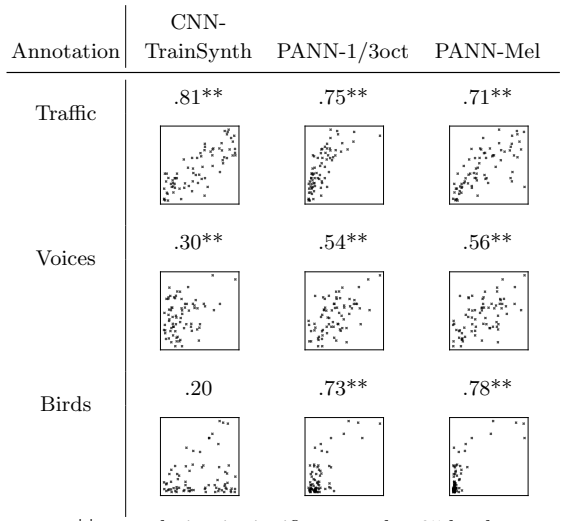
On a GPU V100, PANN-FTO processes 175 seconds of real-time audio within a single second.

2.3.2 Results

The Pearson-r correlation results for the three selected PANN-FTO classes are presented in Figures 2.6a and 2.6b, alongside the correlations for the CNN-TrainSynth and the PANN-Mel models. PANN-FTO demonstrates overall superior performance compared to the CNN-TrainSynth model for the 3 sound sources. It only exhibits slightly lower performance compared to the CNN-TrainSynth model on traffic for GRAFIC, and got similar correlations for birds on Lorient-1k. Furthermore, the performance of PANN-FTO closely rivals that of PANN-Mel, despite PANN-Mel utilizing more refined spectro-temporal representations as input. PANN-FTO appears to outperform PANN-Mel for traffic correlation on Lorient-1k. However, this observation is likely influenced by the low number of data points on which the correlation is calculated.

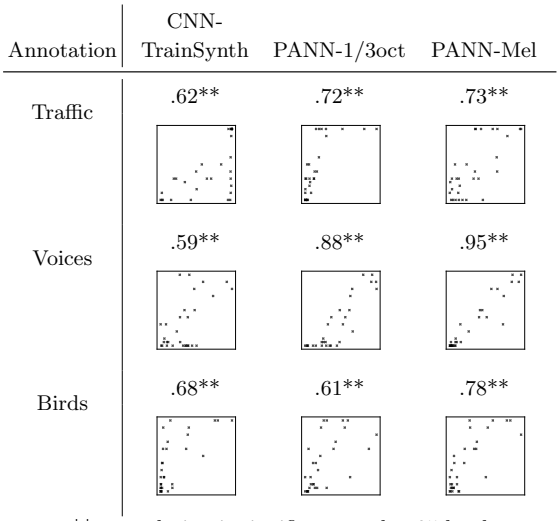
2.4 Qualitative performance analysis

Previously exploited datasets such as Lorient-1k and GRAFIC are relatively small, encompassing less than 4h of audio recordings in total. Unfortunately, no larger-scale audio datasets in the literature include annotations for the time of presence of sound sources. Due to this limitation, we now focus on Lorient Cense, a large-scale dataset of FTO recordings presented in Chapter 1, and that doesn't contain any explicit annotations.



**. correlation is significant at the .05 level
*. correlation is significant at the .01 level

(a) Pearson correlation on GRAFIC (n=74).



**. correlation is significant at the .05 level
*. correlation is significant at the .01 level

(b) Pearson correlation on Lorient-1k (n=30).

Figure 2.6 – Pearson correlations between annotations and predictions for two datasets.

We conduct a qualitative analysis of this dataset to assess if the PANN-FTO prediction results align with our expectations.

2.4.1 Methods

Dataset processing

The entire dataset encompasses over 500k hours of FTO measurements. Given the computational efficiency of PANN-FTO, estimating source levels across the entire dataset would necessitate about 3k hours of computation. Consequently, we opt to sample the dataset within specific time intervals and spatial areas to showcase the efficiency of PANN-FTO in the context of large-scale urban datasets.

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

As detailed in section 2.3.1, the subsets are normalized using the maximum level observed within the dataset of 33,443 recordings collected between January and March 2020, with the top 1% percentile excluded. This established level offset is consistently applied to all other subsets of the Lorient Cense dataset.

Predictions processing

The PANN model predictions are not standardized. For instance, the predictions for the class "Bird Vocalization" in GRAFIC and Lorient-1k, discussed in Section 2.3, fall within the range of 0 to 0.2, while the corresponding annotations for each traffic, voice and birds annotations span from 0 to 1. This annotation scale of 0 to 1 represents a linear and bipolar scale, ranging from "sound source not present at all" to "sound source always present". Similar discrepancies are observed in traffic predictions (0 to 0.4) and voice predictions (0 to 0.6). To address this disparity and ensure alignment with perceived time of presence annotations, we employ linear interpolation without intersection, calculated on the GRAFIC and Lorient-1k predictions. This approach effectively adjusts traffic, voices, and birds predictions to a range consistent with perceived time of presence values, enabling fair inter-class predictions comparisons.

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

2.4.2 Results

Temporal analysis

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

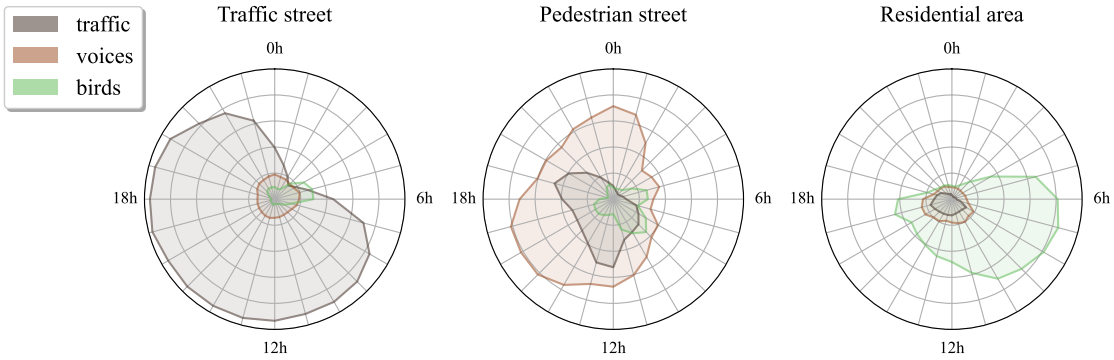


Figure 2.7 – Clock graphs representing the estimated perceived time of presence of traffic, voices and birds, on a sensor close to a traffic street, a pedestrian street and a residential area. The source time of presence prediction is averaged per hour. The time period considered ranges from January 1st to March 1st of 2020, with a random data sampling rate of 14% per day across 67 sensors, resulting in a total of about 500 hours of recording.

We observed highly contrasting distributions of traffic, voices, and birds (see Figure 2.7), all in line with our expectations:

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

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

— Within the residential area, birds are consistently more active throughout the day, with activity peaking about two hours before sunrise. The predictions also show a slight increase at sunset, aligning with expectations. Additionally, we observe minor

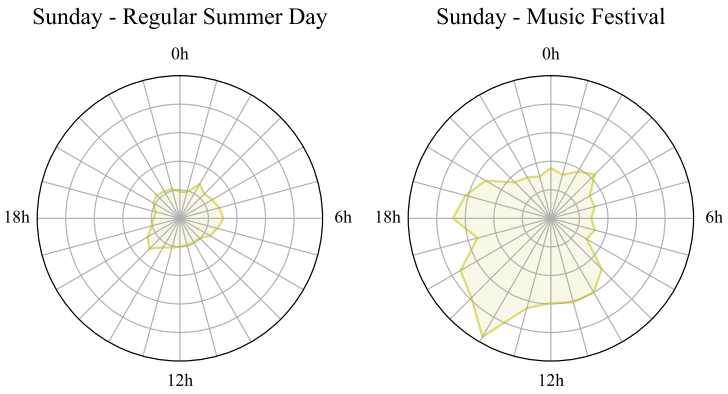


Figure 2.8 – Clock graphs representing the presence of music during a music festival, and on a regular summer day. The source presence likelihood is averaged per hour. The time period considered for Figure a) encompasses every sunday of July 2021, taking into account all data from 5 sensors, resulting in a total of about 200 hours of recording. The time period considered for Figure b) is the 8th of August 2021, taking into account all data from 5 sensors, resulting in a total of about 100 hours of recording.

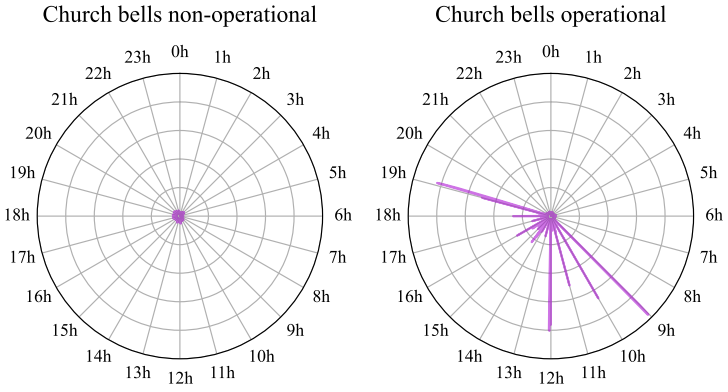


Figure 2.9 – Clock graphs representing the presence of church bells on January 2020 when the bells were operational, and on October 2020, when they were not. The source presence likelihood is averaged per minute. The time period considered for Figure a) is October 2020, taking into account all data from 1 sensor, resulting in a total of about 300 hours of recording. The time period considered for Figure b) is January 2020, taking into account all data from 1 sensor, resulting in a total of about 600 hours of recording.

fluctuations in voice activity between 7h and 8h, and between 17h and 18h, likely linked to people leaving and returning home.

PANN-FTO also demonstrates effectiveness in predicting the presence of other sound sources. Figure 2.8 illustrates a substantial increase in music presence associated with the "music" PANN-FTO class during a festival day compared to a regular sunday. The "Festival Interceltique de Lorient", an outdoor music event, exhibited heightened musical activity on the Sunday 8th of July 2021, primarily in the afternoon. Of particular note is the prominent main outdoor concert that began at 14h30, during which musicians paraded through the streets. The beginning of this parade correlates with the increase of music activity predicted by PANN-FTO in this time period.

Furthermore, Figure 2.9 illustrates the temporal patterns associated with the "church bell" PANN class. During January 2020, when the bells were in full operation, PANN-FTO predictions show that they rang in the opening minutes of each hour. In October 2020, the bells of Notre-Dame-De-Victoire were not ringing due to some structural damages. PANN-1/3 oct has indeed extremely low predictions for the church bell class on this time period. The predictions for church bells are significantly higher at 9h, 12h05, and 19h05. These results are probably due to the Angelus, a daily Catholic prayer traditionally recited three times a day, usually at 7h05, 12h05, and 19h05. In certain churches, the 7h05 prayer may be postponed to a later hour, which seems to be 9h for the Notre-Dame-De-Victoire church.

Spatial analysis

Throughout the entire sensor network, we identify distinct patterns in sound behaviors based on the general sensor location and time periods between January 1, 2020, and March 1, 2020 (see Figure 2.10). Notably, boulevards exhibit significantly higher levels of traffic sounds compared to other street types. Pedestrian streets and areas near shops and bars tend to have a greater prevalence of voices. During nightlife periods, the sensors close to regular shops tend to feature lower levels of voices than the sensors close to bars and nightclubs. Residential areas show an increased presence of birds, coupled with a low presence of voices and traffic, particularly in the morning, aligning with the expected tranquility of nearby parks. Furthermore, our analysis reveals that voice activity predominantly occurs during rush hours and nightlife periods, being notably absent during the early morning. In contrast, traffic sounds are consistently present across all sensors, except during the night hours when the city experiences an overall decrease in activity.

In France, civil defense sirens are activated every first Wednesday of the month precisely at 11h45 These sirens are distributed across various locations in the city. Three

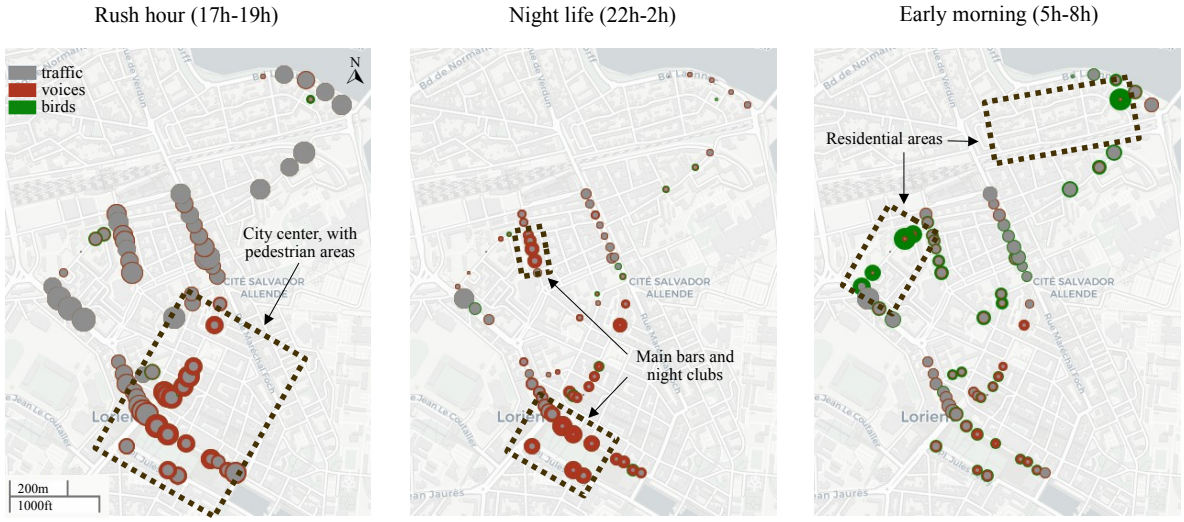


Figure 2.10 – Maps representing the time of presence of traffic, voices and birds over different time intervals. Each node has 3 colors, the area of each color being related to the time of presence of their respective sound source. The circles are not superposed, but should be seen as one circle with the area of each color within it representing the time of presence of each sound source. The time period considered ranges from January 1st to March 1st of 2020, with a random data sampling rate of 0.7% per day across 67 sensors, resulting in a total of about 600 hours of recording.

main sources are situated in the city center of Lorient. Figure 2.11 illustrates the prediction of "Civil Defense Siren" presence for each first Wednesday of the month between January 2020 and July 2020, at 11h45. In comparison to this specific time, the predictions for the presence of sirens for the remainder of the day (when civil defense sirens are not activated) are negligible, being 70 times lower on average. During siren activation, the prediction scores tend to be higher for sensors located closer to the siren sources.

2.5 Discussion

2.5.1 Advantages of using FTO measurements

While it's possible to argue for an alternative approach involving sensors that capture raw audio, make predictions on sound source presence at regular intervals, and send only the predictions to a server to address privacy concerns, there are compelling reasons to consider predicting sound sources directly from FTO measurements. Storing predictions

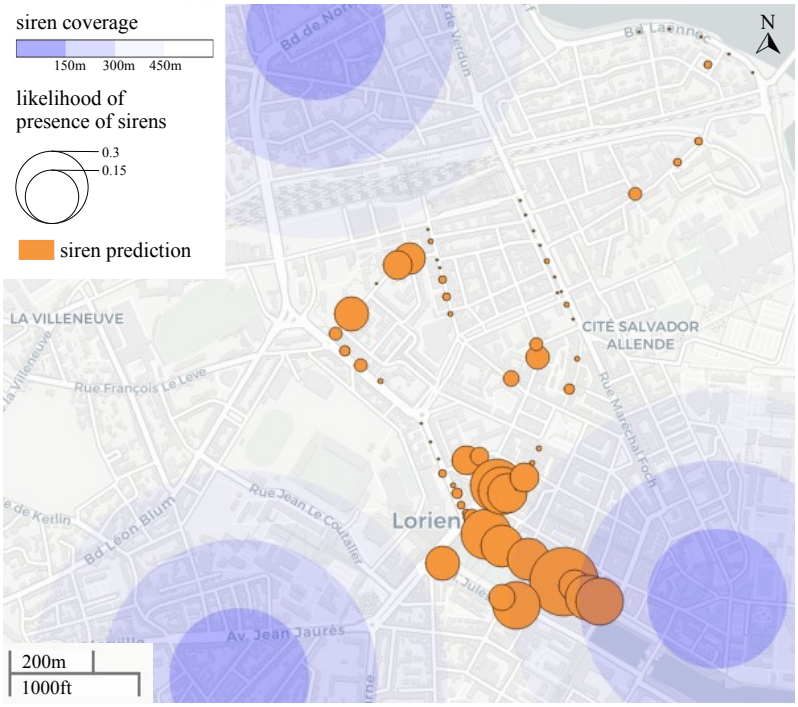


Figure 2.11 – Map illustrating the predicted presence of civil defense sirens on the first Wednesday of each month at 11h45. The map displays the locations of sirens and their estimated coverage. The time period considered encompasses every first wednesday of the month at 11h45 between January and July of 2020, taking into account all data from 67 sensors, resulting in a total of about 5 hours of recording.

rather than FTOs necessitates sensors capable of making such predictions. This implies a requirement for more advanced, potentially higher-cost sensors. Moreover, relying solely on predictions would mean that future access to the original audio data would be lost. In this context, predicting from FTO data with ever improving AI models could potentially lead to more efficient predictions in the future, even considering that we have access to raw audio in this day and age. Furthermore, as FTO measurements from sound level meters are normed (IEC 2013), there are many already recorded datasets where PANN-FTO could be used to make sources predictions (Farrés 2015; Mietlicki et al. 2015; Nilsson et al. 2007; Torija et al. 2013).

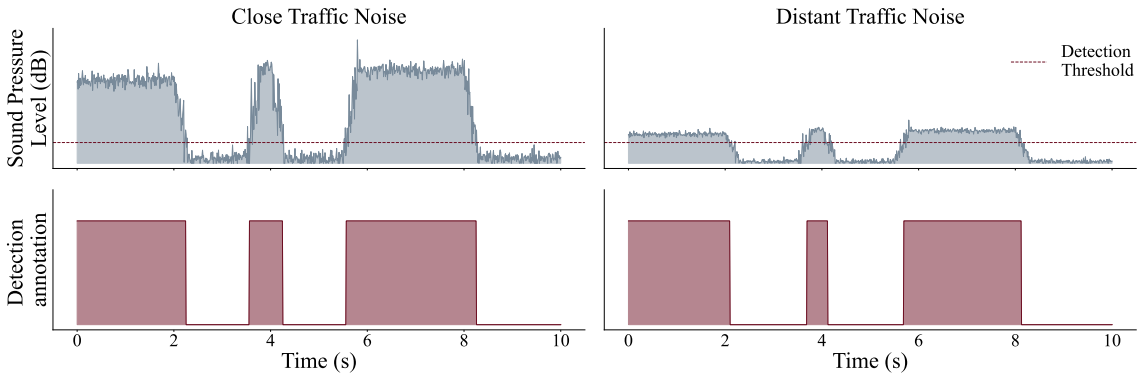


Figure 2.12 – Differences between time of presence annotation procedure, and mean of detection annotation on a synthetic example. On this example, close traffic noise would have a perceived time of presence close to 1/1, and distant traffic noise a perceived time of presence close to 0/1, even if their mean of detection annotation would be extremely close.

2.5.2 Motivations for using perceived time of presence instead of duration or dominance

The evaluation of the presence of sound sources in human assessments is often approached using the concept of "dominance," which aims to capture the perceived level of each sound source (Axelsson et al. 2010; Hong et al. 2015; Mitchell et al. 2021). We believe that the term "dominance" can be meaningful when multiple sound sources are present in a scene. Most studies suggest measuring dominance as a score between 0 and 10. However, in our view, this approach has certain limitations. Implicitly, asking for the "dominance" of a source implies that other sources are being dominated. Therefore, evaluating dominance should involve more of a comparative assessment (e.g., sound source A is twice as dominant as sound source B, which is equally dominant as sound source C), rather than providing individual scores for each sound source (e.g., each sound source has a dominance score between 0 and 10). Consequently, the term "dominance" has been replaced with "time of presence" in some soundscape evaluation studies (Aumond, Can, De Coensel, et al. 2017; Gontier, Lavandier, et al. 2019; Lavandier, Aumond, Can, et al. 2021), a choice also made in this manuscript.

It is crucial to acknowledge that the concept of "time of presence" comes with inherent limitations and should not be confused with the precise determination of the duration of a sound source being active in a given environment, which would require specifying exact

onsets and offsets. The term "time of presence" functions as a straightforward lexical substitution for the perceived overall level of a sound source (Lavandier and Defréville 2006), suggesting a holistic assessment of its presence within a soundscape. Notably, a sound source may receive a maximum time of presence score while being physically present above a certain audible threshold only half of the time, as human perception would naturally fill gaps in auditory information during a holistic evaluation. Similarly, a source physically present and audible for a significant duration, but at a low level or very large distance, would result in a perceived time of presence score close to none. Consequently, datasets annotated with onsets and offsets of sound sources, such as Singa::Pura's dataset (Ooi et al. 2021), are not suitable for evaluating the perceived time of presence. This phenomenon is showcased with a synthetic example on Figure 2.12.

2.5.3 Limitations of acoustic indices compared to AI-based source predictions

While the lack of robustness of acoustic indices was addressed in Section 2.1, we also demonstrated in Section 2.4 that the utilization of PANN-FTO enables studying a more extensive number of sound sources. This is achieved without the need to create a specific acoustic index for each distinct sound source, as PANNs inherently feature 527 output sources. This approach opens up possibilities for a more extensive comprehension of the urban acoustic environment, as exemplified with the music, church bells, and civil defense sirens analysis in Figures 2.8, 2.9, and 2.11.

2.5.4 Potential applications and limitations of PANN-FTO predictions

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

While our results have demonstrated the predictive capabilities of PANN-FTO, extending beyond the primary classes of traffic, voices, and birds, it is important to exercise caution in its application. There is no guarantee of its performance when used on other sound classes for time of presence predictions. Before deploying the model for a new

sound class, a preliminary study is essential to verify the model's compatibility and the correlation of its results with the target sound class annotations.

2.6 Conclusion of the chapter

In this chapter, we have explored a novel approach to accurately predict the time of presence of different sound sources in urban soundscapes. Our approach leverages deep learning models, specifically Pre-trained Audio Neural Networks (PANNs), and a transcoding algorithm to convert FTOs into Mel spectrograms, which are used by PANNs models as input.

Our findings demonstrate that PANNs, whether using Mel spectrograms (PANN-Mel) or transcoded Mel spectrograms from third-octave measurements (PANN-FTO), can accurately predict the time of presence of sound sources in various urban environments, specifically for traffic, voices and birds. This approach provides a robust and efficient means of assessing soundscape quality, especially in situations where relying solely on human-generated annotations is impractical. Additionally, we have shown the versatility of such an approach, which allows to potentially assess the presence of over 500 different sound sources. This capability allows for more in-depth analyses of soundscapes, leading to a better understanding of the acoustic environment in urban areas. However, it remains unproven whether those other PANN classes are similarly correlated with time of presence, as the literature on time of presence assessment has focused on only a limited number of sources.

Takeaways and Perspectives

This chapter provides an answer to our first research question:

RQ1

To what extent can the presence of sound sources be predicted from fast third-octaves recorded by acoustic sensors?

Key Takeaways

— PANN-FTO, a pre-trained classifier adapted to FTO data, can accurately predict the time of presence of traffic, voices, and birds in urban soundscapes.

— These sound source predictions enable meaningful analysis of large-scale sensor networks.

— The approach is versatile and can potentially assess the presence of over 500 sound sources, although the performance of the model for most of these classes has not yet been formally evaluated.

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

Chapter 3

SoundSourceViz: A Visual Analytics Tool for Soundscape-Oriented Urban Planning with Acoustic Sensor Networks

ABSTRACT

This chapter introduces SoundSourceViz, a visual analytics framework designed to support soundscape-oriented urban planning through the exploration of data collected by acoustic sensor networks. SoundSourceViz uses a three-step analytical workflow with interactive multi-source visualizations: (A) polar area diagrams for spatial comparison, (B) geo-referenced 3D data vases for linking spatial location and temporal patterns, and (C) 2D data vases for precise temporal comparisons. The framework allows urban planners to examine spatial and temporal variations in the presence of multiple sound sources, and to compare conditions before and after soundscape design (SSD) interventions. Feedback from domain experts highlights the value of our approach for supporting data-informed decisions in urban sound planning.

Context

The content of this chapter is based on the main-authored scientific publication submitted to Computer Graphics Forum and titled *"SoundSourceViz: A Visual Analytics Tool for Soundscape-Oriented Urban Planning with Acoustic Sensor Networks"*, currently under review.

MAIN CONTRIBUTION

Reference: Modan Tailleur, Vincent Tourre, Pierre Amond, Mathieu Lagrange, and Ophélie Moreau (2025), « SoundSourceViz: A Visual Analytics Tool for Soundscape-Oriented Urban Planning with Acoustic Sensor Networks », *in*: *(submitted to Computer Graphics Forum)*

This contribution builds on our main-authored Internoise article "Towards Better Visualizations of Urban Sound Environments: Insights from Interviews", which informed the selection of a use case scenario for potential visualization strategies. The details of this work are not presented in this manuscript, as it serves solely as supporting material.

SUPPORTING CONTRIBUTION

Reference: Modan Tailleur, Pierre Aumond, Vincent Tourre, and Mathieu Lagrange (2024), « Towards better visualizations of urban sound environments: insights from interviews », *in*: *Internoise*

This paper explores how sound source presence is represented in urban environments, and for whom such representations are most relevant. To this end, eleven 45-minute interviews were conducted with various urban stakeholders. These interviews investigated current usage patterns, the strengths and weaknesses of existing tools, and the relevance of including representations of sound sources.
One of the main use cases for representing sound source presence, highlighted during the interviews, is the management of noise complaints by citizens and industrial stakeholders. For these groups, having access to detailed data at specific times and locations is considered essential. In many cases, even a representation of a single sound source may be sufficient, as the concern often centers on a particular disturbance.

A second important use case concerns providing urban soundscape overviews for non-experts. Citizens who are considering moving to a new neighborhood, for example, might be interested in understanding the typical sound environment of an area. In this context, interpolated maps can provide useful overviews, although several participants raised concerns about interpolation being somewhat misleading. Simpler indicators that are easy to understand were seen as particularly important for this use case. The ability to listen to audio samples was also identified as a valuable addition to visual representations.

A third use case is the methodological support of urban planners. These professionals often work with a wide variety of environmental data, including information about air pollution or soil quality. As a result, their needs for sound-related information are typically specific to each project, and the visualizations must be adapted to the types of data they have collected.

The results from this study are considered as prior knowledge for Chapter 3 of this manuscript.

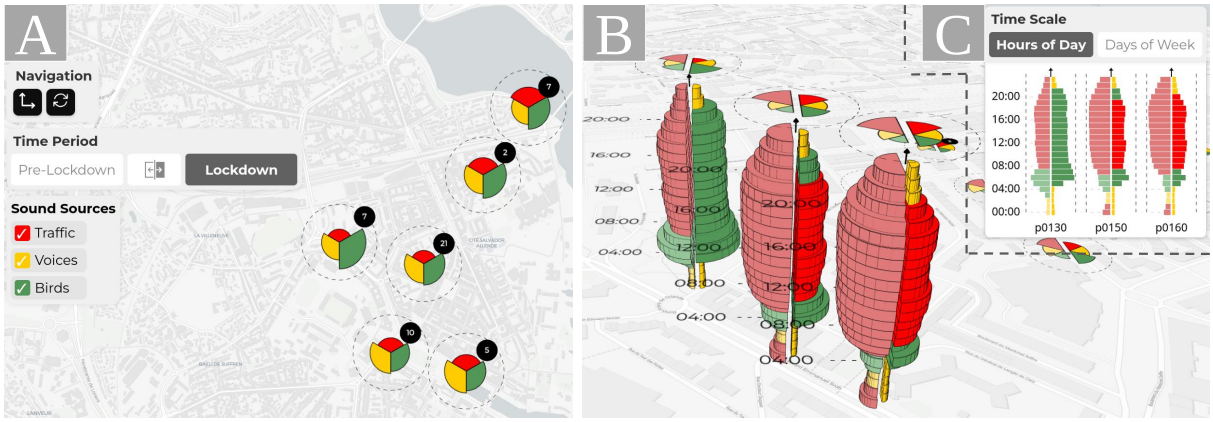


Figure 3.1 – Steps of our soundscape visual analytics framework: (A) The polar area diagram network enables spatial comparison and exploration; (B) The geo-referenced 3D data vases link the spatial and temporal dimensions of the acoustic sensors; (C) The 2D data vases ease the temporal comparisons. The system is shown in comparison mode for steps B and C: the left side of the data vases corresponds to the period before the soundscape design intervention; the right side represents the period after.

3.1 Introduction

Since the sound environment is influenced by urban planning decisions, gaining a better understanding of it can serve two key purposes: evaluating the outcomes of past interventions, and guiding the design of future projects in a sound-informed manner. Noise maps (European Commission 2002) are the only official form of sound visualization currently used by municipal technical departments and urban planners to support sound-informed urban planning decisions. They depict sound levels computed using simulation models that integrate source data (e.g., traffic flow), urban morphology, and standardized propagation algorithms. An example of a noise map is shown in Figure 3.2. Noise maps are typically used to support urban planning strategies which consider sound primarily as a nuisance. In contrast, alternative urban planning approaches known as Soundscape Design (SSD) strategies (Bild et al. 2016) focus not only on mitigating unwanted noise, but also on promoting pleasant sounds through environmental modifications, or acoustic installations (Asdrubali et al. 2014; Hellström et al. 2014; Licitra et al. 2010).

Because SSD prioritizes perceptual outcomes, the evaluation of such interventions often relies heavily on questionnaires designed to capture subjective experiences (Asdrubali et al. 2014; Aspuru et al. 2014; De Coensel et al. 2010). Since questionnaires are labor-intensive and limited to specific times and locations, we proposed in the previous chapter

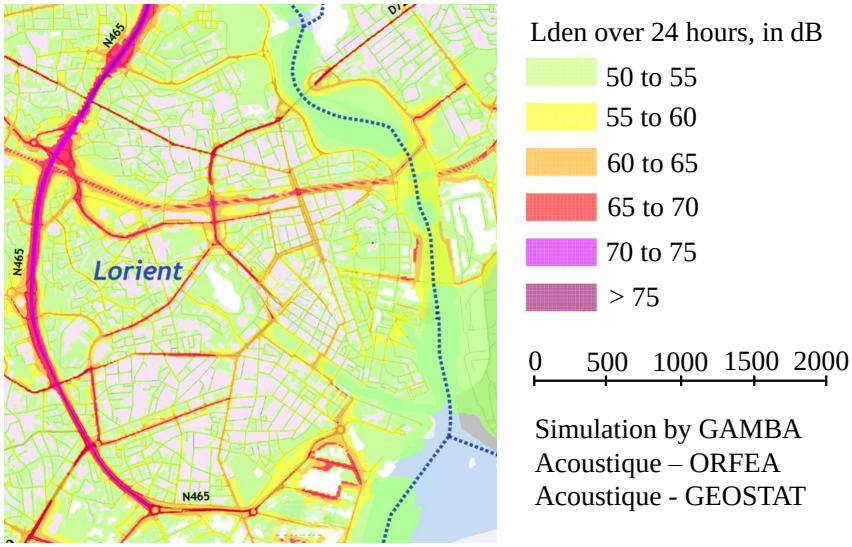


Figure 3.2 – Lorient Noise Map from the 2013 NAP (Ville de Lorient 2013).

a method for estimating perceptual presence scores for traffic, voices, and birds directly from acoustic measurements.

In practice, a sensor is placed in a specific area of interest to gain preliminary insights into the soundscape. Acoustic data is highly localized and its range rarely extends beyond a radius of 50 meters in urban environments (Aumond, Can, Mallet, et al. 2018). SSD interventions at large spatial scales thus run the risk of inaccurate impact assessment. To address this limitation, sensor networks have been developed in various research projects (Can, Picaut, et al. 2021; Dove et al. 2022; Farrés 2015). While these deployments offer richer spatial and temporal coverage, they also introduce new challenges for urban planners. First, the volume and granularity of the data can be overwhelming. Second, interpreting multivariate sound information—such as the presence of traffic, voices, and birds—requires tools capable of helping in the analysis of multi-source patterns that vary across both space and time. Third, comparing soundscape data collected before and after an SSD intervention involves detecting subtle changes between the two time periods.

We believe that addressing these challenges requires a visual analytics approach specifically designed for sound-informed urban planning. Such a tool should help users intuitively explore, compare, and interpret data, in order to go beyond the limitations of traditional noise maps. To this end, we propose SoundSourceViz, a visual analytics framework that provides a sensor-based overview of the soundscape, enable detailed temporal analysis, support navigation between spatial and temporal views, facilitate comparisons between

pre- and post-SSD conditions, and allow flexible filtering of sound sources to suit various analysis needs. The framework follows a three-step analytical workflow supported by interactive visualizations. Our contributions in this article are the following:

— An analysis of the tasks and workflow for soundscape-oriented urban planning: 1) Exploration of a geo-referenced overview, 2) Analysis of temporal patterns, and 3) Comparison of soundscapes,

— The choice of adaptable components to build the visual analytics framework: 1) polar area diagrams to summarize the sound source values over a specific time period, 2) 3D data vases to display efficiently the temporal pattern on the map, and 3) 2D data vases to ease the comparison,

In Section 3.2, we review related work on urban visual analytics frameworks and soundscape visualizations. Section 3.3 presents the design rationales underlying our framework. Section 3.4 describes the various components of the system, while Section 3.5 outlines the analytical workflow it supports. In Section 3.6, we present our evaluation of the framework, followed by a discussion of the findings and future directions in Section 3.7.

3.2 Related Work

3.2.1 Urban visual analytics

Sensors monitoring noise, air quality, or traffic are now widely deployed in cities, producing data that guide planning and policy decisions. As those data can be very dense, visual analytics tools play an important role in helping planners, researchers, and policymakers to analyze this complex information. Geoinformation tools such as ArcGIS (Johnston et al. 2001), and general-purpose platforms such as Urbane (Ferreira, Lage, et al. 2015) and Vis-A-Ware (Ortner et al. 2016) enable the exploration and integration of diverse datasets to support urban planning tasks. In parallel, many visual analytics systems have been developed to focus on specific urban challenges, including air pollution (Deng et al. 2019; Yue et al. 2024), access to sunlight (Miranda, Doraiswamy, Lage, L. Wilson, et al. 2018), transportation networks (Ferreira, Poco, et al. 2013; Itoh et al. 2014), land use (Quercia et al. 2014), and human mobility patterns (Lenormand et al. 2015; Noulas et al. 2012). Other studies investigate the relationships between multiple types of urban data (Chirigati et al. 2016; Malik et al. 2012).

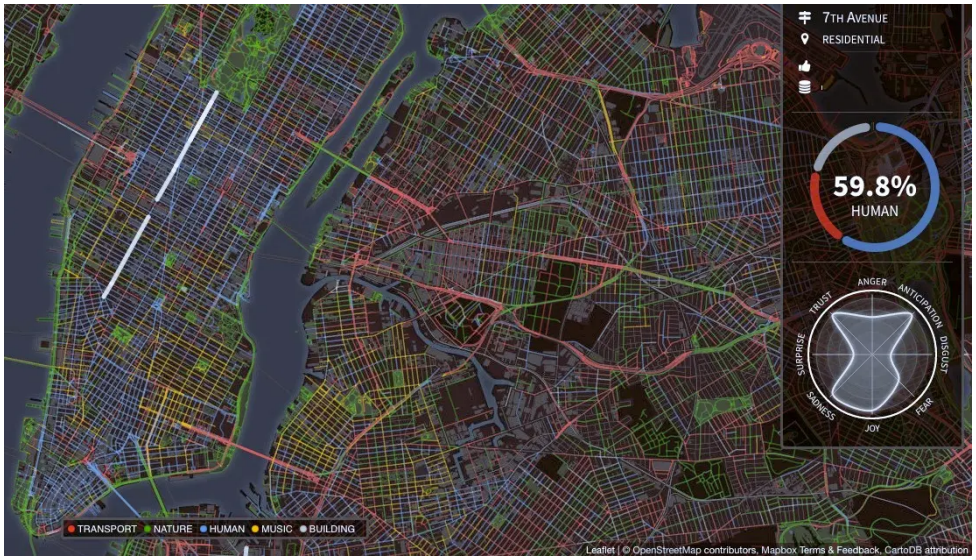
In the field of noise monitoring, Miranda et al. (Miranda, Lage, et al. 2018) introduced a visual analytics tool called Noise Profiler. Noise Profiler is a web-based visual analysis tool that enables interactive exploration of large-scale urban sound data (SPL dBA) collected from a network of acoustic sensors across New York City. It supports analytical queries and real-time monitoring through a visual interface that allows users to examine noise levels by time and location. More closely related to soundscape analysis, Ruff et al. proposed Urban Rhapsody (Rulff et al. 2022), a system designed to support machine learning researchers in exploring large collections of urban sound recordings. Developed in the context of the SONYC (Dove et al. 2022) project, it enables users to interactively query, label, and analyze audio recordings from a network of urban sensors. Our work differs from Urban Rhapsody in that it is specifically designed to support urban planners. In this context, the emphasis shifts from exploring recordings to exploring the soundscape accross space and time at multiple scales, before and after an SSD intervention.

3.2.2 Soundscape Visualization

Aiello et al. (Aiello et al. 2016) leveraged social media data by analyzing image tags from Flickr. They proposed linking sound-related tags to emotional responses and visualized this information on a map using 4 colors to represent dominant sound categories: transport, nature, human, music, and building. Only the most dominant category is shown for each area. When a user clicks on a location, a spider chart is displayed, showing both the emotional profile associated with that area and the relative presence of the five sound categories. This visualization is shown in Figure 3.3(a).

In Moreno et al. (Moreno et al. 2022), audio recordings were collected from 3 different locations over the course of 2 days. The data were manually labeled into several sound categories: Vehicles, Works, City Life, and People. To visualize the temporal evolution of sound at each location, the authors used a space-filling curve, which enabled them to display all 3 locations within a single window. Additional panels provided more detailed information for selected time intervals, allowing for closer analysis of specific time segments of the recordings. This visualization is shown in Figure 3.3(c).

In Lavandier et al. (Lavandier, Aumond, Saul Gomez, et al. 2016), the authors modeled the presence of voices, traffic, and birds using geo-referenced data. Based on these estimates, they derived higher-level perceptual attributes, such as pleasantness. Each of the 3 sound sources is visualized on the map using an icon, with the icon size proportional to the estimated presence of that source. The background of the map conveys the mod-

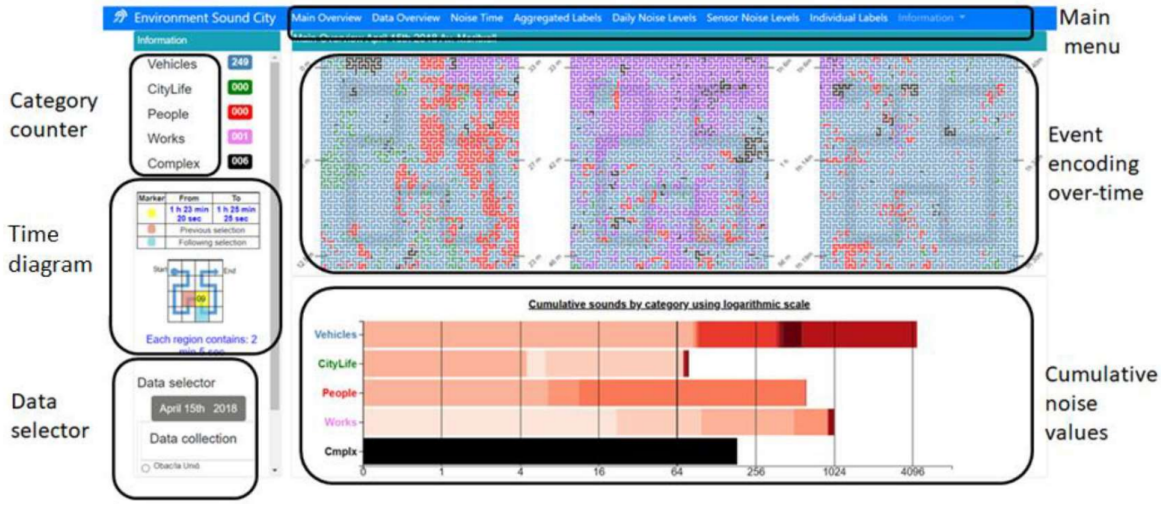


(a) ChattyMaps (Aiello et al. 2016)



(b) Lavandier, Aumond, Saul Gomez, et al. (2016)

Figure 3.3 – Related work in soundscape visualization (1/2).



(c) Moreno et al. (2022)



(d) Lasso platform (Lavandier, Aumond, Saul Gomez, et al. 2016)

Figure 3.3 – Related work in soundscape visualization (2/2).

eled pleasantness, using a color gradient to indicate variations from pleasant to unpleasant areas. This visualization is shown in Figure 3.3(b).

In (Saúl Gomez et al. 2017), the authors proposed new visualizations for representing soundscapes on maps. They introduced custom symbols to depict loudness, using a sound propagation shape that is more or less filled depending on the sound level. A color gradient is used in parallel to convey perceived pleasantness. To represent temporal variations, the symbols are duplicated—showing both daytime and evening sound conditions on the same map.

The Lasso project builds on the work conducted in CartAsur (Lavandier, Aumond, Saul Gomez, et al. 2016). In this visualization tool, users can select a perceptual attribute—such as the presence of traffic, voices, birds, or the pleasantness and eventfulness—via a dropdown menu. The data are collected through questionnaires, with each dot on the map representing a specific survey location. Only one perceptual attribute can be visualized at a time. Each attribute is assigned a unique color, and its level of presence is indicated through a gradient of transparency. When clicking on a dot, additional information appears, including a picture of the location, detailed scores for all perceptual attributes, and options to filter data by specific time periods (e.g., night, day, or evening). To facilitate comparison, the interface includes a duplicated map on the right with synchronized zooming, allowing users to compare different locations, attributes, or time periods side by side. This visualization is shown in Figure 3.3(d).

Those existing soundscape visualization methods are often limited in either spatial or temporal resolution. Indeed, they typically rely on data that is either temporally coarse (e.g., survey-based) or spatially sparse (e.g., localized measurements). In contrast, our work addresses the challenges posed by the visual analysis of multivariate sound source presence data recorded continuously across a dense urban sensor network. This context requires the development of new visualization strategies capable of representing spatial and temporal information while supporting comparison tasks relevant to urban planning.

3.3 Design Rationale

In this section, we first describe the design constraints, then outline the tasks users can perform, followed by the design goals we formulated to support those tasks.

3.3.1 Design constraints

Lorient's noise map (shown in Figure 3.2) reveals substantial differences in noise levels even across the width of a single street. Capturing such spatial variations would require an unrealistically dense sensor deployment, including in private areas, whereas current deployments are typically limited to public infrastructures. The Lorient Cense sensor network (Can, Picaut, et al. 2021) is the densest acoustic sensor network deployed to date, yet the average distance between closest sensors is approximately 50 meters, which is insufficient to capture such spatial variations. Given these limitations, our tasks and design choices focus on sensor-level analysis, treating each sensor as an independent unit without relying on spatial interpolation.

3.3.2 Task Analysis

To support soundscape-oriented urban design, planners need tools that enable them to explore, analyze, and compare soundscape data from several sensors in a way that reflects both spatial and temporal variability. We identify the following key analytical tasks that a soundscape visual analytics framework should support in the context of urban planning:

— **T1: Exploration of a geo-referenced overview of the soundscape.** To interpret the data recorded by the sensors, urban planners must first gain an overview of the soundscape. Via a spatial exploration, the visualization should allow them to quickly identify sensors that are particularly relevant to their analysis. In this context, soundscape is represented by the presence levels of traffic, voices, and birds. As discussed in Chapter 1, these 3 sources provide a meaningful representation of soundscape quality.

— **T2: Analysis of temporal patterns in the soundscape.** Temporal patterns are often implicitly considered in urban design decisions, as noted by (Wunderlich 2023). Differences between night and day, weekdays and weekends, or across seasons have been shown to significantly influence urban soundscapes in various studies (Can, Aumond, et al. 2018; Guo et al. 2022; Hong et al. 2017; J. Liu et al. 2013; Rulff et al. 2022). Urban planners must therefore be able to analyze how the presence of different sound sources changes across these patterns to detect regularities, spot anomalies, and inform their SSD intervention.

— **T3: Comparison of soundscapes before and after SSD**. One of the core analytical tasks of soundscape-oriented urban planning involves evaluating the effects

of an SSD intervention. To do so, urban planners must be able to efficiently compare the soundscape characteristics before and after the intervention. This includes examining changes in sound sources presence, but also assessing how temporal patterns may have shifted (e.g., reductions in traffic noise during work hours, or increased presence of natural sounds at specific times of day).

3.3.3 Design Goals

To guide our design process, we derive several goals from the analytical tasks presented in the previous section.

— **G1: Provide a sensor-based soundscape overview of the network.** The framework should provide an overview of the soundscape using data from each sensor's actual location. As explained Section 3.3.1, sensors are typically too far apart for simple spatial aggregation to yield meaningful or reliable results. Instead of using spatial aggregation between sensors, the framework should display each sensor location, enabling users to explore the sensor network spatially and access multivariate information related to the presence of different sound sources (T1).

— **G2: Enable temporal pattern analysis for selected sensors.** Users should be able to select specific sensors of interest and view detailed temporal patterns (daily, weekly, monthly, or yearly profiles) for the presence of the three sound sources (T2).

— **G3: Provide a clear link between spatial and temporal view.** Given the importance of temporal structure in understanding soundscapes (T2), the framework should provide geo-referenced links between spatial glyph overview and detailed temporal patterns.

— **G4: Support comparisons between pre- and post SSD.** The framework should support visualizing pre- and post SSD periods, and facilitate comparison through glyph and temporal pattern designs specifically adapted for this purpose (T3).

— **G5: Ensure flexibility in the number of sound sources displayed.** Planners may not always need to consider all sound sources simultaneously. The framework should allow them to filter or focus on a subset of sources, while maintaining the coherence and interpretability of the components (T1, T2, T3).

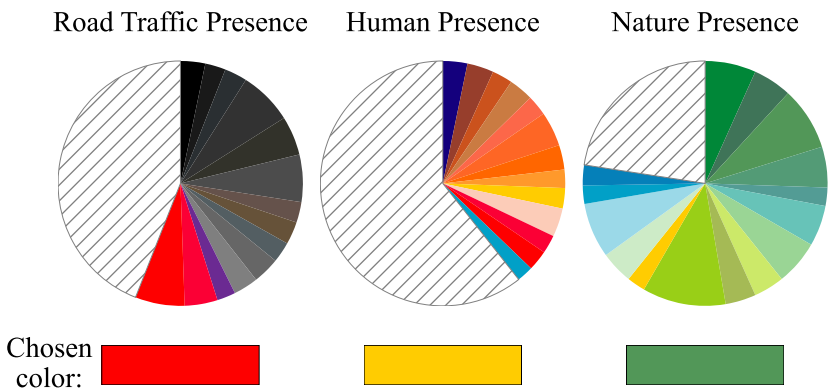


Figure 3.4 – Colors related to the presence of each sound source, along with the final color chosen for our study.. Colors that received less than 2% of answers are not represented here, and are merged into the hashed area.

3.4 Components

In this section, we introduce the components of SoundSourceViz, our visual analytics tool designed to support soundscape-oriented urban planning using data from an acoustic sensor network. We conceptualize the dataset from the sensor network as a form of space-time cube (Kraak et al. 2005), where sound source presence is encoded across time along the vertical axis of the map. To describe the role of each component within this spatiotemporal framework, we adopt the terminology proposed by Bach et al. (Bach et al. 2017).

3.4.1 Color scheme used in components for traffic, voices, and birds

To represent traffic, voices, and birds, we assign a distinct color to each sound source. This color selection is common to all the design goals (G1, G2, G3, G4, G5).

Our color selection is informed by experimental data from Lavandier et al. (Lavandier, Aumond, Saul Gomez, et al. 2016), who conducted an online survey in December 2015 with 150 participants. Respondents were asked to choose 3 colors from a palette to represent the presence of (1) human activity, (2) nature, and (3) road traffic. The aggregated responses are summarized in Figure 3.4. These survey results were collected as part of the study but were not analyzed or used in the original publication.

From the analysis of the responses, we aim to select a combination of three colors that

meets three key criteria:

— Basemap Contrast: Colors should not be too close to gray, ensuring clear visibility against standard gray-scale basemaps.

— Within-Category Consistency: A color should not deviate too far from the average color associated with its category.

— Perceptual Separation: The colors should be sufficiently distinct to avoid visual confusion.

We begin by enforcing the basemap contrast criterion, filtering out any triplets that include colors too close to gray hues. We use a minimum distance threshold of 0.06 units in CIE 1976 UCS color space (Robertson 1977), based on the findings of Wilson and Crawford (D. L. Wilson et al. 1989), who showed that this distance threshold marks the point at which symbols become reliably distinguishable from background colors in a fast recognition task. Letting D_{CIE} denote the Euclidian distance function in CIE 1976 UCS space, this constraint can be formally expressed as:

$$\mathcal{S} = \Big\{ \mathbf{s} \mid \forall\, \mathrm{Col} \in \mathbf{s}, \quad \forall\, g \in [0, 255],\ D_{CIE}(\mathrm{Col}, (g, g, g)) > 0.06 \Big\} \tag{3.1}$$

To enforce within-category consistency, we then compute, for each remaining color of each triplet, the weighted standard deviation of its RGB Euclidian distance (D_{RGB}) to the category's average color. We define an outlier as any color whose distance to the mean exceeds one standard deviation. For example, in the case of human presence, this approach tends to retain orange, brown, and yellow hues, while excluding blue hues that fall outside the dominant color range.

Among the valid triplets $\mathcal{S}$, we ensure perceptual separation by selecting the set $\mathbf{s}^*$ that maximizes the minimum pairwise distance between the three selected colors—ensuring they are as distinguishable as possible:

$$\mathbf{s}^* = \arg\max_{\mathbf{s} \in \mathcal{S}} \left(\min_{\substack{\mathrm{Col}_i, \mathrm{Col}_j \in \mathbf{s} \\ i<j}} D_{RGB}(\mathrm{Col}_i, \mathrm{Col}_j) \right) \tag{3.2}$$

These calculations yield the following color assignments: a red shade (R: 254, G: 0, B: 0) for traffic presence, a yellow shade (R: 255, G: 204, B: 1) for human presence, and a green shade (R: 82, G: 151, B: 88) for nature presence, as shown in Figure 3.4.

This selection aligns with how similar sound sources have been represented in previous studies. Puyana-Romero et al. (Puyana-Romero et al. 2019) observed that traffic sounds

are often depicted using gray tones. Social sounds are typically associated with blue (e.g., busy pedestrian streets) or lighter colors such as white and yellow (e.g., playgrounds). Natural sounds are commonly linked to green hues in park settings and blue tones near water. Green color can have restorative psychological effects (Berman et al. 2008), evoke a sense of safety (Pravossoudovitch et al. 2014), and support creativity-enhancing environments (Lichtenfeld et al. 2012). These cognitive and emotional effects further support the use of green color to represent natural soundscapes.

3.4.2 Glyph design for spatial exploration

To provide a broad overview of the urban soundscape in its spatial dimension, we adopt a glyph design that supports the identification of points of interest across the city by encoding the average presence levels of 3 sound sources, across the entire time period that is represented (G1). Each glyph represents one measurement point. According to the terminology proposed by Bach et al. (Bach et al. 2017), this representation performs a *time flattening* operation on the space-time cube, aggregating temporal information into a spatial summary.

We propose the use of polar area diagrams, as illustrated in Figure 3.5(c), derived from Nightingale's rose diagrams (Nightingale 1857). Our diagrams consist of three colored slices, each corresponding to one sound source. The area of each slice is proportional to the source's level of presence. According to Li et al. (Li et al. 2010), symbol size is perceived as the second most dominant visual channel after color. This makes polar area diagrams particularly well suited to our use case: color distinguishes the sound sources, while area communicates their relative levels of presence. The design is also adaptable to varying numbers of sound sources (G5), as shown in Figure 3.6. A dashed circle indicates the maximum possible radius, serving as a visual reference for maximum presence.

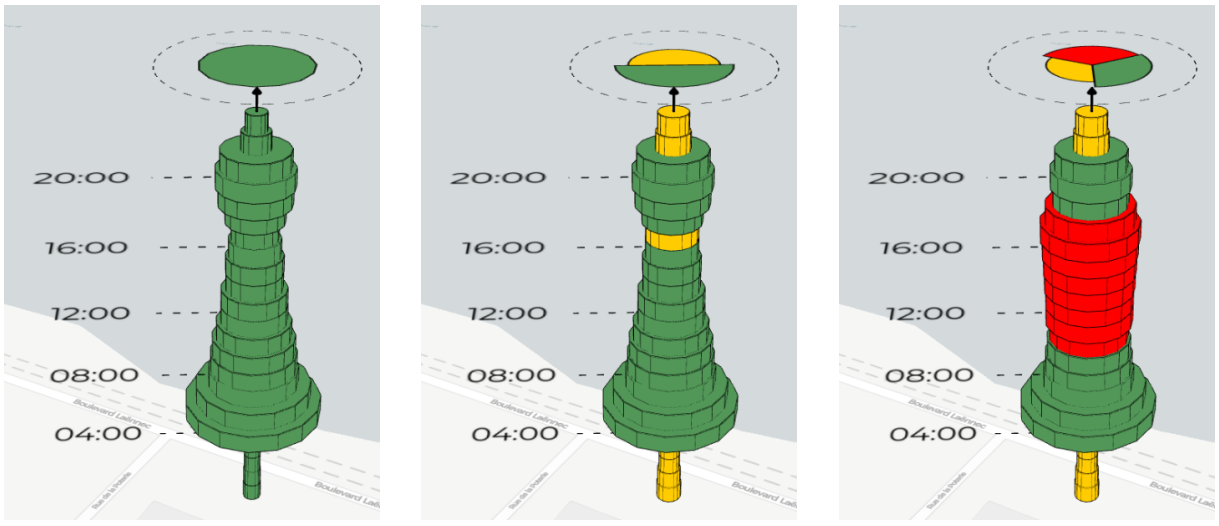
We define the area of a polar area diagram as the area of this circle represented with a dashed line. This area is thus identical for all sensors. Following the recommendations of (McNabb and Laramee 2019; McNabb, Laramee, and M. Wilson 2018), we set the area of each polar area diagram to occupy 2.5% of the available screen space in pixels, independantly of the zoom level. However, this design could lead to significant visual clutter at low zoom levels, as sensors are located close to one another. To address this, we apply a glyph aggregation method that groups overlapping polar area diagrams. As discussed in Section 3.3.1, sensor data can't be meaningfully averaged and represented with a single glyph, as sensors may be too spatially distant relative to the fine-grained variability of the

acoustic environment. Instead, we adopt a greedy aggregation algorithm that iteratively removes overlapping polar area diagrams based on their pairwise spatial distance. The details of this algorithm are provided in Appendix A. As shown in Figure 3.1(A), the number of polar area diagrams hidden beneath the one displayed at a given zoom level is displayed in the top-right corner of the visible diagrams.



(a) Pre-SSD (b) Comparison (c) Post-SSD

Figure 3.5 – Polar area diagrams showing traffic (red), voices (yellow), and birds (green), across the 3 different period visualization modes.



(a) 1 source displayed (b) 2 sources displayed (c) 3 sources displayed

Figure 3.6 – Polar area diagrams and data vases, with 1, 2, or 3 sources displayed.

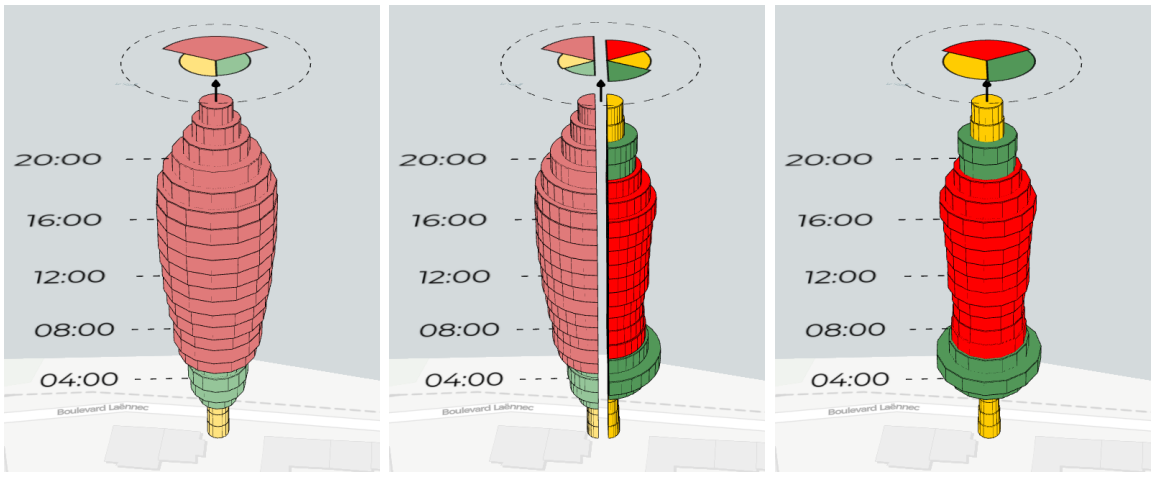
3.4.3 Data vases for geo-referenced temporal exploration

To visualize temporal patterns across multiple locations while preserving spatial anchoring (G2, G3), we propose to extend the 3D data vase concept introduced by Thakur et al. (Thakur et al. 2010). This type of 3D representation has been notably used by Gautier et al. (Gautier et al. 2021) to illustrate the spatio-temporal evolution of the COVID-19 pandemic. Whereas previous data vases typically encoded only a single variable along the vertical dimension, our version supports multivariate encoding. Specifically, color is used to distinguish our 3 distinct sound sources, and the vertical axis represents time, segmented into discrete intervals (e.g., hours of a day, days of a week). Each horizontal slice of the data vase corresponds to one time interval. Its color indicates the most dominant sound source (i.e., the one with the highest presence score), and its radius encodes the magnitude of that dominant source during the interval. Displaying only the dominant sound source for each interval simplifies visual comparison across data vases. It produces a distinctive silhouette that makes characteristic temporal patterns easier to recognize. In the case of daily patterns, illustrated in Figure 3.7(c), each data vase is divided into 24 sections, each representing one hour of the day. Dashed lines, along with time labels, extend from the data vase to provide temporal references. The design of the data vase is also adaptable to the display of a varying numbers of sound sources (G5), as shown in Figure 3.6. Following the terminology of Bach et al. (Bach et al. 2017), this representation corresponds to a *repeated drilling* operation on the space-time cube, followed by a *3D rendering*.

These temporal patterns (e.g., daily or weekly) are derived by averaging the presence values of the 3 sound sources across those time intervals. As discussed in Section 3.3.1, due to spatial heterogeneity, averaging values across different sensors is not feasable. However, averaging over time for a single sensor remains valid and meaningful, as it reflects recurring temporal structures in the sound environment at that specific location.

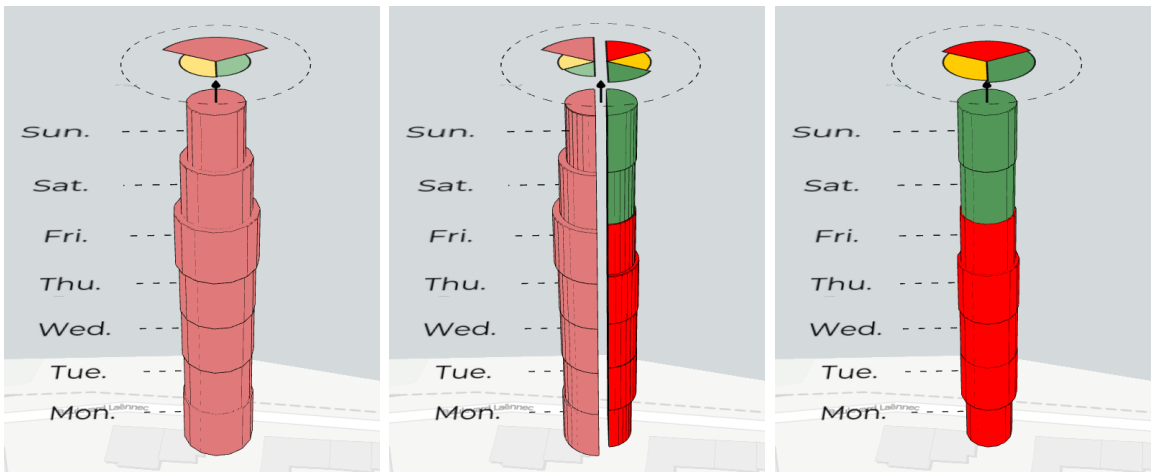
The polar area diagram remains visible at the top of each vase, maintaining visual continuity with the previous step (G3). Its dashed outer circle also offers a rough reference for the maximum magnitude of presence of each sound source.

Daily pattern



(a) Pre-SSD (b) Comparison (c) Post-SSD

Weekly pattern



(d) Pre-SSD (e) Comparison (f) Post-SSD

Figure 3.7 – Data vases showing traffic (red), voices (yellow), and birds (green) presence for two temporal patterns, across the different period visualization modes.

3.4.4 2D data vases of data vases for precise temporal exploration

The 2D projection of the 3D data vases, shown in Figure 3.1(C), offers a clearer view of temporal patterns (G2) by removing the spatial perspective present in the 3D vase. To remain consistent with the patterns presented in Section 3.4.3, this view also displays the

most dominant sound source for each time interval. This representation can be described as a *2D rendering* of the *repeated drilling* operation on the space-time cube, as defined by Bach et al. (Bach et al. 2017). The 2D data vase includes dashed vertical lines that indicate the maximum possible value for each section, serving as visual reference points to facilitate magnitude comparison. Unlike the previous components presented in Sections 3.4.2 and 3.4.3, this visualization is not geo-referenced; only the sensor's identifier is displayed. This 2D data vase aims to reduce the potential visual clutter associated with 3D representations and facilitate temporal comparisons across sensors that are spatially distant on the map.

3.4.5 Support for pre- and post-SSD comparisons: color hues and component adaptations

To support comparisons between two time periods (G4), we provide visual encodings for both pre- and post-SSD data, using distinct color hues for each period and adapting components to support a dedicated comparison mode.

The pre-SSD glyphs and data vases are represented with the same color hues but with a 25% reduction in lightness in the HSL (Hue, Saturation, Lightness) color space. This approach, illustrated in Figures 3.5(a) and 3.7(a), aligns with Brewer's recommendations for sequential color schemes (Brewer 1999), as variations in lightness is often used to convey ordered data. In our case, the pre- and post-SSD conditions represent a temporal sequence, making this a perceptually appropriate choice.

Users can activate a comparison mode in which each component is adapted to simultaneously display both time periods (pre- and post-SSD) using a mirrored design. The left side of the component represents the pre-SSD period, and the right side represents the post-SSD period. Examples of these adaptations to the original components are shown in Figure 3.5(b) and Figure 3.7(b). For glyphs, the design requires further modification, as the original circular layout can't be directly mirrored. Instead, all sound sources are visualized on a half-circle. The core visual encodings remain the same: color encodes the sound source, area indicates presence level, and the dashed outer circle represents the maximum possible presence.

This mirrored design enables direct visual comparison without duplicating glyphs. Ondov et al. (Ondov et al. 2018) found that mirrored layouts perform comparably to duplicated ones for detecting correlation or identifying maximum differences between paired

diagrams. Their compact design makes them well suited for map-based displays with limited screen space. Ondov et al. (Ondov et al. 2018) also found that mirrored layouts enhance user performance in comparison tasks compared to traditional side-by-side views for bar charts. Given the vertical, bar-like structure of the data vases, this layout is also particularly well suited to highlight differences between the two time periods for this component.

3.5 Analytical workflow

In this section, we present the 3-step analytical workflow supported by the component designs described in Section 3.4. This workflow progressively guides the user from a spatial overview to detailed temporal patterns. It is not intended to be strictly linear or prescriptive, as users may move fluidly between the steps as new insights emerge. Red reference numbers in Figure 3.9 indicate the identifiers of the buttons proposed on SoundSourceViz. These identifiers are used throughout the following sections to describe the buttons in each step of the analysis. The workflow is summarized in Figure 3.8.

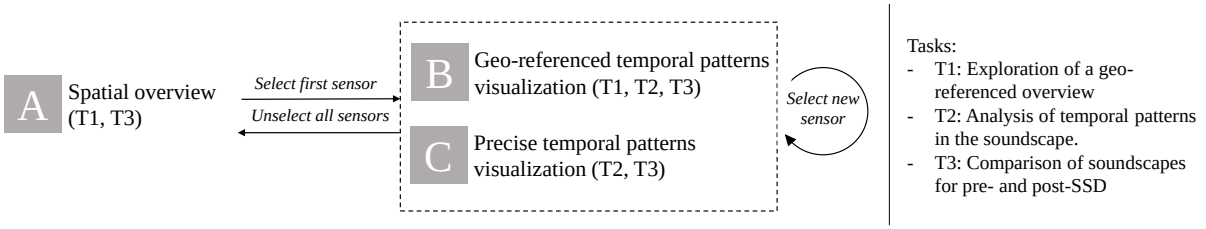


Figure 3.8 – Analytical workflow of SoundSourceViz. The steps of the workflow, A, B, C, correspond to those presented in Figure 3.1.

3.5.1 Step A: Geo-referenced, glyph-based overview of the soundscape.

In Step A, the user primarily performs task T1 (Exploration of a geo-referenced overview of the soundscape), but also task T3 (Comparison of soundscapes before and after SSD). The step is presented in Figure 3.1(A).

During Step A, the user can explore the soundscape spatially through a network of geo-referenced glyphs displayed on a background map, which provide an overview of the presence of the sound sources on the monitored area (G1). Because glyphs are rendered

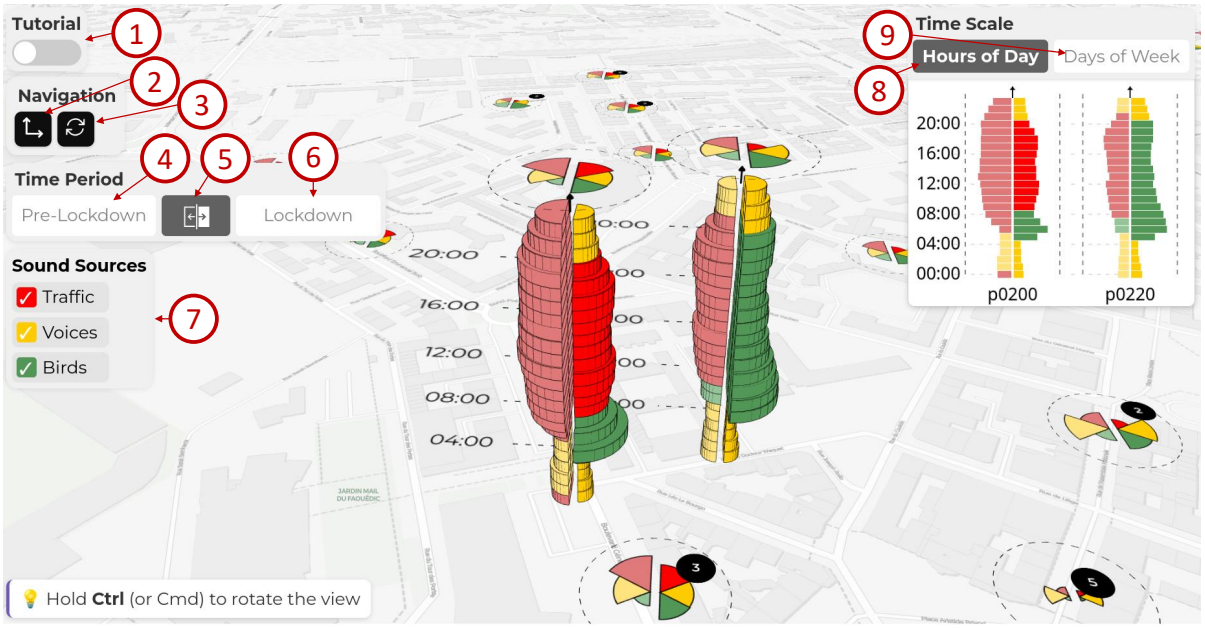


Figure 3.9 – Overview of the visual analytics framework as used during the expert analysis. For clarity, button sizes have been enlarged. Red circles indicate the identifiers of interaction buttons and are included solely for reference in the paper.

with a fixed pixel width, they are aggregated at low zoom levels, with the method described in Section 3.4.2. Users can pan using dragging interaction to explore the map and zoom into specific areas of interest using the mouse wheel to reveal more glyphs. As explained in Section 3.4.1, the background map color palette is a gray scale to avoid any confusion with the value of interest.

To enable exploration of the soundscape across both time periods and facilitate their comparison, users can toggle between pre-SSD (Button 4) and post-SSD (Button 6) to focus on a specific time period and compare value between sensors (G1), or activate a dedicated comparison mode (Button 5) using the controls in the "Time Period" section of the interface to be able to perform the pre- and post-SSD comparison (G5). Users can dynamically select or deselect sound sources using the source selection checkboxes (Button 7), depending on their specific interests or planning priorities (G5). To aid interpretation, hovering over a glyph displays a tooltip showing the corresponding sensor identifier and the associated time period (pre- or post-SSD).

3.5.2 Step B: Geo-referenced temporal patterns visualization for the analysis of selected locations.

In this step, users can perform all three tasks: T1 (Exploration of a geo-referenced overview of the soundscape), T2 (Analysis of temporal patterns in the soundscape), and T3 (Comparison of soundscapes before and after the SSD intervention). The primary objective of this step is to fulfill design goal G3, which aims to establish a link between the spatial and temporal views (Steps A and C). It is presented in Figure 3.1(B).

After identifying sensors of interest on the map, users can click on their glyphs to display the corresponding 3D data vases (G2). The users can adjust the pitch angle of the 3D view by holding the `Ctrl` key while dragging, and thus visualize temporal patterns (G3). Up to three vases can be displayed simultaneously; selecting a fourth removes the first one. This limit is based on Miller's law (Miller 1956), which suggests that users can effectively process around 7 (± 2) items at a time. In our interface, each data vase in comparison mode presents 2 sets of information (pre- and post-SSD). When 3 vases are shown, this results in a total of 6 elements, which is bellow the cognitive threshold suggested by Miller's law. Clicking on a vase that is already displayed will conveniently remove it.

As with the glyphs described in Section 3.5.1, each vase can be toggled between pre-SSD (Button 4), post-SSD (Button 6), and comparison (Button 5) views, depending on the analysis focus (G4). Sound sources within the vase can also be selected or deselected using the source checkboxes (Button 7). This enables users to focus on specific sources or to reveal the presence of other sound sources by hiding dominant ones (G5). Two navigation buttons are provided for this step: one resets the view to a 2D perspective by setting the pitch angle to zero (Button 2); the other resets both the view and all visualization parameters to their initial state (Button 3). The full legend of the 3D data vases could be difficult to read if displayed permanently, as it creates visual clutter and overwhelms the display with information. To mitigate this, the time references are only partially displayed, as illustrated in Figure 3.7. For more detailed information, users can hover over a segment of a vase to display a tooltip with the sensor identifier, the time period, and the corresponding time interval (e.g. hour of the day) (G3).

3.5.3 Step C: Precise temporal patterns visualization for the analysis of selected locations.

This step supports Task T2 (Analysis of temporal patterns in the soundscape) and Task T3 (Comparison of soundscapes before and after the soundscape design intervention). It is presented in Figure 3.1(C).

When at least one data vase is displayed, a window containing the temporal data of the selection appears (G2). This interaction aligns with the visual information-seeking principle introduced by Shneiderman (Shneiderman 2003): "overview first, zoom and filter, then details on demand" (G2). Additional controls (Buttons 8 and 9) appear as well, allowing users to toggle between daily average data (Figures 3.7(a-c) and weekly average data (Figures 3.7(d-f)). These controls are presented alongside the 2D data vases.

3.6 Evaluation

In this section, we first provide implementation details of our framework. Then, we present the dataset collected by the Lorient CENSE sensor network and explain how we preprocess this data to infer the presence of traffic, voices, and birds. We also describe how we use this observational dataset to simulate SSD-related analysis. We then present the protocol followed during our expert evaluation study, including a guided walkthrough of the interface and a structured feedback evaluation. Finally, we present the results of this evaluation.

3.6.1 Implementation details

Our visualization framework is implemented as a client-side web application. We employ Deck.gl (v9.0.0) (Wang 2019) for WebGL-based rendering of spatial data, integrated with MapLibre GL (v3.6.0) for interactive basemap visualization. The system runs entirely in the browser without requiring a backend server, ensuring portability and ease of deployment. [1]

1. Source code available at: `https://github.com/modantailleur/paperSoundSourceViz`

3.6.2 Lorient CENSE dataset processing

We use data from the sensor network deployed in the city of Lorient, France (Can, Picaut, et al. 2021), and presented in Chapter 1. For this study, we focus on data recorded between January 1st, 2020 and May 11th, 2020. From the sensors active during this period, we exclude those that recorded less than two weeks of data. After filtering, 52 sensors remain, providing a total of approximately 150,000 hours of spectral data. We use PANN-FTO model introduced in Chapter 3 to estimate the presence of traffic, voices, and birds from the acoustic data. The predictions are performed on 10-second windows and require approximately 500 hours of computation. Because the network provides continuous monitoring, we are able to capture rich temporal variations in the soundscape, such as daily and weekly patterns. However, due to the decline in sensor activity after mid-2020, we are not able to reliably analyze seasonal trends.

Although the dataset does not contain labeled examples of actual SSD implementations, we can still study changes in the urban soundscape in response to known external events. In this paper, we compare soundscape data across periods without (pre-SSD) and with (post-SSD) COVID-19 lockdown restrictions, as well as between weekdays (pre-SSD) and Saturdays (post-SSD). The pre-lockdown period spans from January 1st to March 16th, 2020, while the lockdown period begins on May 11th, 2020. While these scenarios were not explicitly designed as SSD interventions, they serve as proxies for typical SSD-related analysis.

3.6.3 Experts panel

We evaluate our visual analytics framework with 3 experts: 2 technical officers responsible for noise-related matters in French cities (Subject A and Subject B), and one urban planner (Subject C).

Subject A is involved in environmental and noise-related issues for a French city of 50,000 inhabitants. He/She contributes to environmental health initiatives, including projects on noise exposure. Strategic noise maps are occasionally used in his/her work, particularly when urban development projects raise concerns about noise impacts (e.g., in the context of building permits). He/She also takes part in inter-municipal meetings to review preliminary results from strategic noise mapping that feed into the Noise Action Plan (NAP).

Subject B works for a metropolitan authority of 700,000 inhabitants, on air quality and

noise related matters within the Directorate for Public Health and Ecological Transition. His/Her current responsibilities focus heavily on regulatory noise-related work, including the update of strategic noise maps and the development of the NAP.

Subject C is an urban planner currently involved in the development of large-scale planning documents, such as intermunicipal land-use plans and territorial coherence schemes. He/She supports local elected officials in identifying the needs of their territories and formalizing development strategies. Subject C has collaborated on planning cases where acoustic constraints were relevant, particularly in the context of modifying land-use plans near major roadways.

3.6.4 Guided Analysis

The only widely adopted sound-related visualization is the noise map, as presented in Section 3.1. The visual analytics tool we propose offers a novel approach, going beyond what is currently available to urban planners and municipal technical services. To help participants become familiar with our framework, we designed a guided analysis, available by toggling the "Tutorial" button (Button 1). This tutorial presents explanatory text and interaction prompts, followed by five questions. The first set of questions introduces the polar area diagrams, guiding users in switching between the two time periods (pre- and post-SSD) and activating the comparison glyph (T1, T3). Another group of questions serves a similar purpose but focuses on the data vase visualizations (T2, T3). The final question presents a more open-ended exploration task, asking users to examine a specific area of the map and respond to a prompt that simulates a realistic urban planning scenario (T1, T2, T3). All the questions referenced above are explicitly listed in Figure 3.10. [2]

3.6.5 Evaluation protocol

After completing the guided analysis, experts evaluate the visual analytics framework on a laptop equipped with a ChiMei InnoLux 0x1406 display (13.9-inch diagonal, 1920×1080 resolution).

Participants first rate their level of agreement with the statement:

"This action helped me better understand the sound environment and how it evolved between the two time periods."

2. Guided analysis is available at: https://modantailleur.github.io/paperSoundSourceViz/, by toggling the "Tutorial" button. The original french version is available by toggling the "Fr" button.



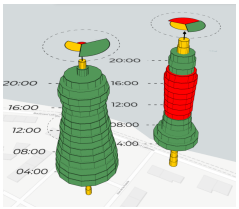
Q1: *"Among the two displayed symbols, which one shows a higher presence of birds?"*

Interactions allowed: *none*



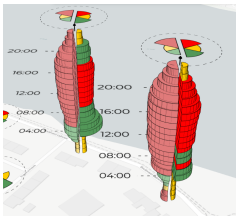
Q2: *"On which of the two sensors did the presence level of birds change the most between before and during lockdown?"*

Interactions allowed: *period button, sound sources checkboxes*



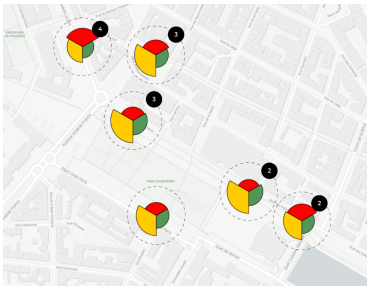
Q3: *"Among the two sensors, which one shows more voices between 1pm and 8pm? Does the presence of voices appear to be different depending on the day of the week?"*

Interactions allowed: *period button, sound sources checkboxes, pitch rotation*



Q4: *"On which of the two sensors did the presence of birds change the most between before and during lockdown, between 16:00 and 20:00?"*

Interactions allowed: *period button, sound sources checkboxes, pitch rotation, time scale button*



Q5: *"While exploring this part of the map and comparing several sensors, find a sensor that seems particularly more lively at night on Saturday than during the week."*

Interactions allowed: *period button, sound sources checkboxes, pitch rotation, time scale button, click, zoom, navigation button*

Figure 3.10 – Questions for the guided analysis. Questions were asked in French but are translated here into English. Each figure illustrates a different task involving interpretation of sound presence across time, space, and source.

This statement is evaluated for each of the following actions, using a 5-point Likert scale (1: Strongly disagree, 2: Disagree, 3: Neither agree nor disagree, 4: Agree, 5: Strongly agree):

— Reading polar area diagrams

— Comparing polar area diagrams

— Reading 3D data vases

— Comparing 3D data vases

— Interpreting 2D data vases

— Comparing 2D data vases

— Exploring the map

Participants are also invited to evaluate the following statement using the same Likert scale:

"This visualization enables a more in-depth analysis than what a standard noise map provides."

The term "standard noise map" refers to the noise map commonly used in urban planning and produced during NAPs, as described in Section 3.2.

Finally, they are encouraged to comment on the strengths and limitations of the visual analytics framework and to provide any additional feedback.

At each step, discussions are conducted informally to better understand their reasoning. The questionnaire results can't be evaluated quantitatively. Instead, our analysis primarily focuses on the content of the discussions, with questionnaire responses occasionally used to support specific points.

3.6.6 Results

The participants indicated that the selected sound sources (traffic, voices, and birds) are largely aligned with established practices, noting their similarity to classifications they have previously encountered, such as mechanical, anthropogenic, and natural sources.

Evaluation of Task 1: Exploration of a geo-referenced overview of the soundscape.

Subjects A, B and C found the polar area diagrams to be a clear and effective way to visualize the presence of sound sources (rated 4/5, 5/5 and 4/5 respectively). Sub-

ject A initially found the diagrams somewhat confusing due to their resemblance to pie charts, but explained that the interpretation became straightforward after a short period of familiarization.

Participants are highly satisfied with the exploration (rated 5/5 by all). Subject B noted, however, that navigating the map to identify points of interest was initially challenging, particularly when data vases were displayed. Subject C also found navigation somewhat difficult at the city-wide scale. He/She identified the glyph aggregation strategy as the main limitation. He/She emphasized, however, that these issues are specific to large scales and did not arise at neighborhood or local levels, where fewer sensors are shown and glyph overlap is not a concern.

Evaluation of Task 2: Analysis of temporal patterns in the soundscape.

Subject B stated that being able to see what lies beneath a sensor in three dimensions is particularly interesting. He/She found the visualization of temporal patterns especially valuable. In their current workflow, they typically produce one noise map per type of time interval, since the simulation tools used to generate these maps are not available for every possible interval. He/she appreciated that the data vases visualization enables the exploration of multiple time intervals within a single view.

Subject A and B noted that while the geo-referenced 3D data vases are helpful for recognizing general temporal trends (rated 4/5 by both), the 2D data vases are more effective for retrieving precise information (rated 5/5 by both). Subject C also preferred the 2D data vases (rated 4/5) to the 3D data vases (rated 3/5) for analyzing the soundscape, as he/she noted difficulties in reading the 3D data vases due to the visual cutter, and perspective issues it implies.

Evaluation of Task 3: Comparison of soundscapes before and after SSD

Subject B mentioned that he/she used the comparison mode only occasionally, preferring to alternate manually between the pre- and post-SSD views. In contrast, Subject A relied exclusively on the comparison mode to analyze differences between the two time periods and rarely switched between individual pre- and post-SSD views. Subject C mentioned that the fact of being able to compare pre-SSD and post SSD are one of the main strenghts of the tool.

3.7 Discussion

3.7.1 Improving the spatial exploration

Subject C suggested that spatial interpolation of source presence scores could improve the spatial exploration of sensor data. However, generating a smooth, continuous soundscape map through basic interpolation between sensors could be misleading, since acoustic environments often show strong local variations, as explained in Section 3.3.1. In this context, although using glyphs at individual sensor locations imposes limitations on large-scale spatial exploration, it remains the most reliable and interpretable method available for visualizing data collected by acoustic sensor networks. Upon discussing these limitations, Subject C proposed, as a potential direction for future work, to reveal hidden glyphs on mouse hover to better support exploration. He/She also suggested that adding more contextual cues, such as a more details on the base map, would help users in deciding where to focus their attention during exploration.

3.7.2 Improving temporal analysis

Subject B suggested an improvement to the temporal analysis by proposing the ability to hide specific sections of the column, allowing users to focus on selected time intervals.

Alternatives in 2D Data Vases Design

When clicking on a sensor, its 3D data vase appears (Figure 3.1(B), with a 2D data vase displayed in a separate window (Figure 3.1(C)). In our framework, the 2D data vase follows the same visual encoding as the 3D version: only the color of the most dominant sound source is shown for each time interval. An alternative design could involve displaying the individual presence curves for all 3 sound sources. However, we chose to maintain consistency with the 3D data vases, as it supports a smoother transition between Step B and Step C of the analysis workflow illustrated in Figure 3.1.

Pattern recognition

In this work, we propose that users visually compare different time patterns (e.g., daily, weekly) using the framework. However, prior research by Miranda et al. (Miranda, Doraiswamy, Lage, Zhao, et al. 2016) has shown promising results in the visualization of automated pattern recognition related to sound. In future versions of our framework,

users could double-click on a sensor to automatically highlight other sensors that exhibit similar temporal patterns.

3.7.3 Framework generalization

The framework can be extended to support additional temporal patterns, such as yearly, seasonal, or monthly patterns. It also supports a variable number of sound sources, both in the polar area charts described in Section 3.4.2 and in the data vases introduced in Sections 3.4.3 and 3.4.4. This flexibility suggests that this tool could be repurposed for other types of sensor networks. For example, Subject C proposed using it to visualize the presence of bird species in eco-acoustics research, as part of biodiversity requalification projects. More broadly, the framework could be applied to other forms of urban data. Potential applications outside of acoustics include supporting the analysis of the temporal evolution of airborne pollutants such as NOx, or the distribution of allergenic pollen types.

3.7.4 Extension of the Framework

This work focused exclusively on the visual analysis of data collected from acoustic sensor networks. However, soundscape assessments often include subjective evaluations gathered through questionnaires. These typically capture a broader range of information, such as emotional responses, and the context in which sounds are experienced. In this regard, Subject C suggested that future work could explore integrating these complementary data sources into the visual analytics framework. For instance, aggregated qualitative feedback, such as word clouds representing reported feelings, could be displayed alongside sensor data to enrich the understanding of the soundscape.

3.8 Conclusion of the chapter

In this chapter, we introduced SoundSourceViz, a visual analytics framework designed to support soundscape-oriented urban planning through the analysis of data collected by acoustic sensor networks. Rather than visualizing raw acoustic data, we leverage sound source classification models to estimate the perceptual presence of 3 key sound sources—traffic, voices, and birds—which serve as proxies for mechanical, social, and natural components of the urban soundscape. The framework enables planners to examine the spatial and temporal dynamics of presence of these sources and to compare their evolution

before and after soundscape design (SSD) interventions. SoundSourceViz 3-step analytical workflow supports both high-level overviews and detailed temporal pattern analyses.

The results of our expert evaluation demonstrate that the tool enables a rich and nuanced analysis of sensor network data. Participants particularly valued the ability to compare pre- and post-intervention conditions and to visualize temporal patterns of sound source presence. These findings underscore the potential of data-driven soundscape visualization to support and inform urban design and planning decisions.

Future work could explore several improvements of the framework, such as integrating perceptual feedback (e.g., citizens survey responses) into the visual analysis, or refining the background map to provide more effective contextual cues for interpreting sensor data.

Takeaways and Perspectives

This chapter provides an answer to our second research question:

RQ2

How can a visual analytics tool support soundscape-informed urban planning using acoustic sensor data?

Key Takeaways

- We propose a workflow for soundscape-oriented urban planning: 1) Exploration of a geo-referenced overview, 2) Analysis of temporal patterns, and 3) Comparison of soundscapes,
- We use adaptable visual components to build the visual analytics framework: 1) polar area diagrams to summarize the sound source values over a specific time period, 2) 3D data vases to display efficiently the temporal pattern on the map, and 3) 2D data vases to ease the comparison,
- Feedback from 3 experts demonstrate that the tool enables a rich and nuanced analysis of sensor network data.

The framework proposed in this chapter relies exclusively on visual representations of perceptual attributes. While this visual approach provides valuable insights for urban planners, it remains inherently indirect in conveying the actual acoustic experience. In the following chapter, we introduce a new method that enables direct exploration of the sensor data through listening.

Chapter 4

Audio Skim Generation for Environmental Audio Recordings

ABSTRACT

Listening to audio recordings offers valuable insights into soundscapes by conveying information more directly than soundscape indicators, which need interpretation with acoustic knowledge. However, accessing the rich information contained in environmental audio recordings is often impractical due to their length, making them difficult to explore or analyze by listening. To address this, we introduce an automatic method for generating audio summaries, which we call audio skims, designed to highlight key characteristics of a full-length audio. Our approach balances two objectives: faithfulness, which reflects how closely the summary preserves the most present sound sources and the temporal structure of the full-length audio, and highlightness, which reflects how well the summary represents the most salient sound sources of the full-length audio. We define a controllable parameter z to modulate this trade-off and propose a two-stage audio skim generation pipeline: Main Events Identification (MEI) via K-means clustering (MacQueen 1967) of audio embeddings, and Sequencing of Main Events (SME) using a greedy algorithm optimized for the z-weighted balance between highlightness and faithfulness. We test our method on a dataset of several urban recordings by generating 1-min summaries from 24h duration full-length audio. Through both objective and subjective evaluations, we show that our method produces high-quality audio skims and enables controllable trade-offs between faithfulness and highlightness. This approach supports practical and customizable access to environmental recordings, with applications in soundscape research, urban planning, and public communication.

Context

The content of this chapter is based on the main-authored scientific publication in preparation for submission to Transaction on Audio, Speech, and Language Processing (TASLP) and titled "Audio Skim Generation for Environmental Audio Recordings".

MAIN CONTRIBUTION

Reference: Modan Tailleur, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2025a), « Audio Skim Generation for Environmental Audio Recordings », *in*: *(to be submitted)*

The model developed in the main-authored article "Enforcing Speech Content Privacy in Environmental Sound Recordings using Segment-wise Waveform Reversal" is employed in this chapter to ensure speech content privacy within the audio recordings. This privacy preservation is particularly important, as the audio recordings may be accessed for perceptual evaluation by external participants as well as by the authors themselves. By applying this method, speech content is effectively rendered unintelligible, thereby ensuring a form of privacy enforcement.

SUPPORTING CONTRIBUTION

Reference: Modan Tailleur, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2025b), *Enforcing Speech Content Privacy in Environmental Sound Recordings using Segment-wise Waveform Reversal*, DOI: 10.48550/arXiv.2507.08412

In this paper, we introduce a method that renders speech unintelligible while preserving both the integrity of the acoustic scene, and the overall audio quality. Our approach involves reversing waveform segments to distort speech content. This process is enhanced through a voice activity detection and speech separation pipeline, which allows for more precise targeting of speech.
In order to demonstrate the effectiveness of the proposed approach, we consider a three-part evaluation protocol that assesses: 1) speech intelligibility using Word Error Rate (WER), 2) sound sources detectability using Sound source Classification Accuracy-Drop (SCAD) from a widely used pre-trained model, and 3) audio quality

using the Fréchet Audio Distance (FAD), computed with our reference dataset that contains unaltered speech. Experiments on this simulated evaluation dataset, which consists of linear mixtures of speech and environmental sound scenes, show that our method achieves satisfactory speech intelligibility reduction (97.9% WER), minimal degradation of the sound sources detectability (2.7% SCAD), and high perceptual quality (FAD of 1.40). An ablation study further highlights the contribution of each component of the pipeline. We also show that incorporating random splicing to our speech content privacy enforcement method can enhance the algorithm's robustness to attempt to recover the clean speech, at a slight cost of audio quality.
This approach is used in this chapter to preserve the content privacy of audio recordings in long-term audio recordings. It enables audio playback for expert evaluation, even when the recordings may contain potentially sensitive content.

4.1 Introduction

In (Tailleur, Aumond, Tourre, et al. 2024), interviews with various city stakeholders revealed that acoustic measurements and even high-level perceptual attributes, such as pleasantness or the presence of specific sound sources, are not always intuitive for understanding the quality of a sound scene. Findings suggest that providing short, representative audio summaries of a location could offer a more intuitive and insightful way to convey the essence of the recorded soundscape. In the remainder of this chapter, we refer to the original recording as the **full-length** audio.

According to Schafer (Schaffer 1979), a soundscape can be described through three key elements:

— the keynote, referring to sounds heard frequently enough to form a background against which other sounds are perceived;

— the signal, which designates foreground sounds that attract attention;

— the soundmark, a sound characteristic of a particular location and time, often carrying special meaning for that place.

Keynotes are the sounds most prominently experienced by individuals in real time. This effect is typically captured in cognitive science through experience sampling methods, where individuals report their feelings at various moments throughout the day. Findings in (Ellison et al. 2020) suggest that experience sampling is particularly useful for capturing both general aspects of the experience and its temporal dynamics, which aligns closely with the role of keynotes in a soundscape context.

By contrast, signals and soundmarks tend to play a central role in *a posteriori* evaluations of an experience. In cognitive science, this effect is typically measured through retrospective report methods, where individuals are asked to provide an overall evaluation after an experience has ended. In such evaluations, people emphasize the most salient or intense events (Alaybek et al. 2022), which in the context of soundscapes correspond precisely to signals and soundmarks.

This duality in the assessment of an individual's experience within a soundscape motivates the definition of two complementary objectives for audio summarization. On the one hand, we aim to represent the most present background sounds at different time periods, while preserving their temporal evolution (e.g., wind at night, morning birds, daytime traffic hum). We loosely term this notion **faithfulness**. On the other hand, we also aim to capture the main salient sources, and the ones that define the identity of the origi-

nal recorded place and time. Sounds such as church bells, passing trains, or distinctive sounds specific to a given soundscape fall into this category. We loosely term this notion **highlightness**. Inevitably, these two objectives are in tension. A faithful summary, by emphasizing the most present background sounds and their gradual evolution, tends to smooth out rare but salient events; conversely, a highlightful summary, by showcasing distinctive and memorable sounds, risks losing the background sound patterns, as well as their dynamics and temporal continuity, thereby reducing faithfulness.

In (Oldoni et al. 2015), the authors proposed extracting short, unordered samples from environmental audio recordings using a self-organized map. The extraction can be saliency-based (most noticeable events), frequency-based (most common events), or a combination of both. A similar approach is employed in surveillance video summarization, where the goal is often to extract keyframes or short clips corresponding to salient events (Song et al. 2016; Thomas et al. 2017; Zhang et al. 2016). Since this method produces unordered salient samples, we consider that it primarily supports highlightness. While the frequency-based variant of (Oldoni et al. 2015) reflects a certain form of faithfulness, its lack of attention to chronology and overall dynamics prevents it from fully meeting our definition of faithfulness.

In the video domain, a video skim is typically defined as a concatenated collection of video segments (Apostolidis, Adamantidou, et al. 2021). This concept can be extended to the audio domain by summarizing full-length recordings into concatenated collections of audio segments. Because an audio skim preserves chronology through its ordered sequence, it can potentially capture the dynamics of the original recording and thus align more closely with our definition of faithfulness than an unordered collection of events. Audio skim generation is thus the approach adopted in this chapter. The task of generating audio skims with dual objectives of faithfulness and highlightness is novel, and no prior state-of-the-art methods exist specifically for this problem. Nevertheless, several related works address similar tasks and help contextualize our approach.

Recent approaches to video skim generation rely predominantly on deep learning methods. Both supervised (Argaw et al. 2024; W. Xu et al. 2024) and unsupervised (Apostolidis, Adamantidou, et al. 2020; Apostolidis, Metsai, et al. 2019; He et al. 2019; Jung et al. 2019; Mahasseni et al. 2017; Yuan et al. 2019; Zang et al. 2023) strategies have been explored. The studies employ a wide range of architectures, including transformers, Long Short-Term Memory (LSTM) networks, and Generative Adversarial Networks (GANs). They are typically trained and evaluated on hundreds of hours of video recordings or

movies, often using human annotations or human-produced video skims as supervision or ground truth.

Music summarization aims to produce a shorter version of a track that preserves both its musical structure and content diversity (Cooper et al. 2003; Kim et al. 2006; Shao et al. 2005; C. Xu et al. 2004). A closely related task, known as music thumbnailing, focuses on extracting the single most representative segment of a piece (Burges et al. 2008; Cooper et al. 2002; Jiang et al. 2014; Lee et al. 2016; Logan et al. 2000; Muller et al. 2012; Nieuwenhuijsen 2019). Both approaches typically rely on the detection of repeated structures such as introductions, verses, and choruses, which guide the summarization process and its evaluation.

In written language processing, various approaches have emerged with the goal of enhancing the flexibility and robustness of text-to-text summarization systems. Recent research have more specifically addressed challenges such as generating summaries from incrementally growing text streams (X. Chen et al. 2024), handling low-resource scenarios (Y.-S. Chen et al. 2022), improving the coherence of long summaries (Q. Liu et al. 2021), and effectively aggregating information from multiple sources (Wei et al. 2021). Audio summarization has also been explored in the context of audio speech content (Furui et al. 2004; Vartakavi et al. 2021). In these works, automatic speech recognition (ASR) is used to transcribe the audio, followed by text-based summarization using transformer models. The final audio summary is then reconstructed by mapping the summarized text back to its corresponding positions in the original audio.

Our task differs from the ones described above in several key aspects. First, full-length environmental audio recordings are not engineered like musical pieces and therefore lack an inherent structure that can be exploited for generating a summary. Second, the duration of environmental sound recordings are typically much longer, which makes summarization significantly more challenging. Among the related tasks, video skim generation is arguably the most similar, as it also involves selecting diverse and engaging excerpts. However, its primary goal is to maximize attractiveness, an objective that lies outside the scope of our focus on faithfulness and highlightness. Regarding summarization methods, supervised deep learning approaches are not applicable for our task, as no reference audio skims exist for training or evaluation. Furthermore, the combined objectives of faithfulness and highlightness necessitate methods specifically designed to optimize them.

We believe that a good audio skim is a trade-off between faithfulness and highlightness. To model this trade-off, we introduce a user-defined control parameter z, where

$z = 1$ produces audio skims that emphasize highlightness over faithfulness, and $z = 0$ prioritizes faithfulness over highlightness. Based on this parameter, we propose a method that generates an audio skim adapted to the desired balance. The method consists of two stages: Main Events Identification (MEI) and Sequencing of Main Events (SEI). In the MEI stage, the full-length audio is divided into fixed time periods, and a set number of representative elements are selected per period using K-means clustering (MacQueen 1967). The SEI stage then arranges these elements into a final sequence using a greedy algorithm that accounts for the input value z.[1]

In Section 4.2, we present our method. Section 4.3 details the experimental protocol, Section 4.4 reports the results, and Section 4.5 provides discussions on these findings.

4.2 Methods

4.2.1 General framework

We propose a two-step approach to address the audio skim generation problem, illustrated in Figure 4.1. In the first step, the audio is quantized into a sequence of neural embeddings. In the second step, we optimize a z-dependent objective function to select N embedding indices. Since each embedding corresponds to an audio segment, we retrieve the selected segments and reconstruct the audio skim using overlap-add. We refer to our method for audio skim generation as **SkimEAR** (*Skim generation for Environmental Audio Recordings*).

Step1: Main Events Identification (MEI)

The Main Events Identification (MEI) step, involves 1) segmenting the full-length audio into T consecutive, non-overlapping time periods, 2) extracting k_t representative elements from each period t based on audio embeddings representation, and 3) computing a scarcity score for each element. This procedure is detailed in Section 4.2.2.

Step2: Sequencing of Main Events (SME)

The Sequencing of Main Events (SME) step consists of selecting one audio segment from each of N different time periods among the T periods identified in the MEI step, thereby forming a sequence with N distinct elements. The sequencing algorithm needs to

1. Open source code and audio examples available at: https://modantailleur.github.io/paperAudioSummary/

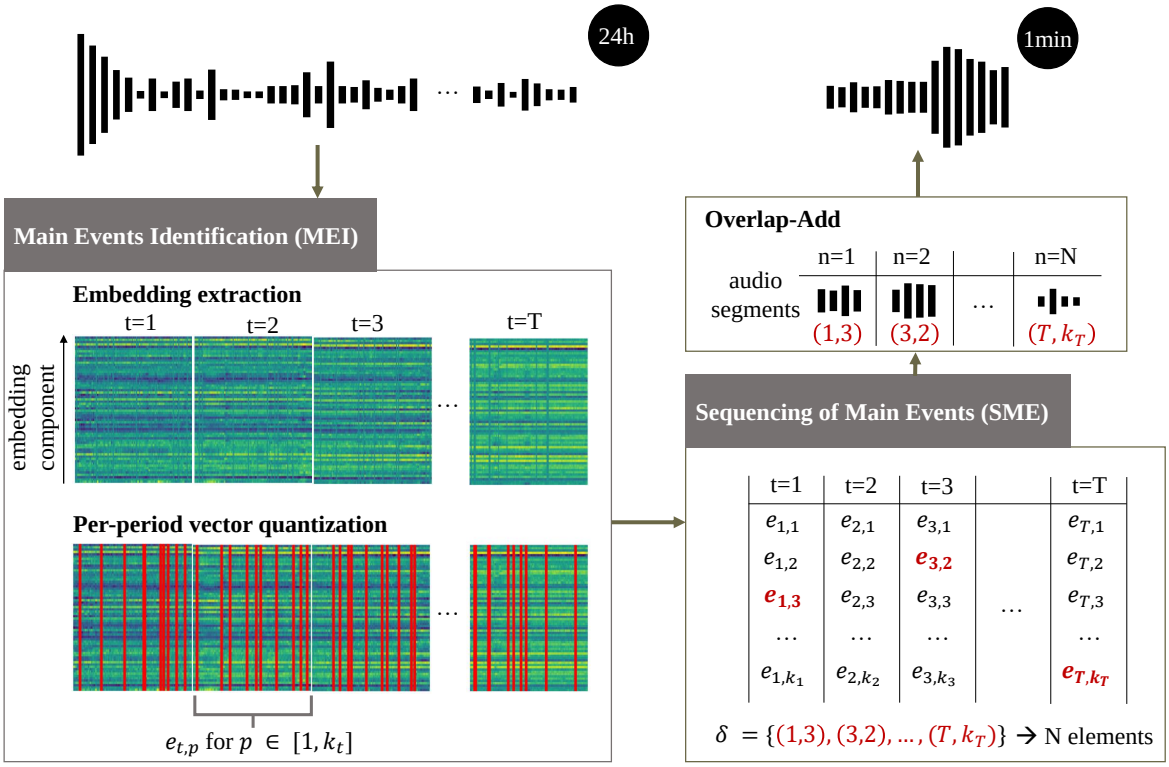


Figure 4.1 – Pipeline for Automatic Audio Skim generation. $e_{t,p}$ represents the pth embedding for period t, selected by a Kmeans algorithm for which we take the medoid of each cluster. δ represents the indices of the audio segments of the sequence.

compose a selection that best balances faithfulness and highlightness, depending on the trade-off parameter z. Achieving this balance requires considering both the time period and the scarcity score assigned to each identified element during the MEI step. Our optimization objective is to find the sequence that maximizes the sum of pairwise distances between all its elements. Since this objective naturally favors highlightness, it is adjusted when higher faithfulness is desired. The detailed formulation of this objective and the proposed optimization strategy are presented in Section 4.2.3.

4.2.2 Main Events Identification (MEI)

For the MEI stage, we first split the full-length audio into T consecutive, non-overlapping time periods. Each period is then treated independently in the remainder of the MEI step.

Next, we extract audio embeddings from each time period of the recordings. To do so, the audio is split into non-overlapping 8-second segments within each period. Specifically,

we compute L-CLAP embeddings, obtained from the CLAP (Contrastive Language–Audio Pretraining) architecture, trained by LAION (Wu et al. 2023). Their best-performing model—referred to as 630k-audioset-best—was trained specifically for environmental audio tasks. It uses HTS-AT (K. Chen et al. 2022) as the audio encoder and RoBERTa (Y. Liu et al. 2019) as the text encoder, on approximately 10,000 hours of audio data.

We then perform K-means clustering (Arthur et al. 2006) on the audio embeddings within each time period t, with clustering carried out independently for each period. The K-means algorithm minimizes the within-cluster variance by solving the following optimization problem:

$$\arg\min_{C,\{\mu_k\}_{k=1}^{K}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2, \tag{4.1}$$

where $C = \{C_1, \ldots, C_K\}$ is the set of clusters, μ_k is the centroid of cluster C_k, and x_i are the data points (in our case, audio embeddings) assigned to each cluster.

Finally, we compute a scarcity score for each 8-s audio segment. For each cluster p in each time period t, the medoid—i.e., the embedding element closest to the cluster centroid—is selected and denoted as $e_{t,p}$. This step involves two controllable parameters: the number of time periods T and the Clustering Ratio (CR), defined as the number of clusters k_t relative to the number of embedding elements within a given time period t. To reflect how frequent each audio segment is within a given time period of the full-length audio, we assign a scarcity score to each element based on the density of its associated cluster. Let $ne_{t,p}$ denote the number of elements in the cluster to which the pth element of period t belongs. The scarcity $s_{t,p}$ is computed as:

$$s_{t,p} = \frac{\max\limits_{0 \le j \le k_t} (ne_{t,j})}{ne_{t,p}} \tag{4.2}$$

The scarcity score $s_{t,p}$ ranges from 1, indicating elements belonging to the densest clusters and thus the least scarce, up to $\max\limits_{0 \le j \le k_t} n_{t,j}$.

4.2.3 Sequencing of Main Events (SME)

Optimisation objective

We aim to balance two conflicting objectives: highlightness and faithfulness. Rather than treating these objectives separately, our approach formulates the optimization problem around highlightness, operationalized through the maximization of sound source di-

versity. To incorporate faithfulness into the optimization objective, we introduce a z-dependent weighting function. As z approaches zero, the function increasingly favors the selection of audio segments that are highly present within each time period and whose temporal positions within the sequence better reflect the structure of the original recording, thereby shifting the objective toward faithfulness.

We define Ω as the set of all possible sequences that form an audio skim, where each sequence consists of the period and cluster indices of one embedding element per time period from a total of N time periods:

$$\Omega = \left\{ \delta = [(t_i, p_i)]_{i=1}^{N} \;\middle|\; \begin{array}{l} t_i \in [1, T], \quad p_i \in [1, k_{t_i}], \\ \\ t_i \neq t_j \quad \forall\, i \neq j, \end{array} \right\}$$

Maximizing highlightness can be seen as finding a sequence that results in the maximum possible pairwise distance between all its elements in the embedding space. The optimization objective is then formulated as follows:

$$\delta_{opt}(z) = \underset{\delta \in \Omega}{\operatorname{argmax}} \sum_{\substack{\{(t_i,p_i),(t_j,p_j)\} \\ \subset \delta, i \neq j}} \omega_z(s_{t_i,p_i}, s_{t_j,p_j}) d(e_{t_i,p_i}, e_{t_j,p_j}) \tag{4.3}$$

To incorporate faithfulness into the optimization, we introduce a weighting function, denoted as ω_z, defined as:

$$\omega_z(s_{t_i,p_i}, s_{t_j,p_j}) = \frac{\theta_z(|i-j|)\beta_z(s_{t_i,p_i})\beta_z(s_{t_j,p_j})}{\sum_{\substack{(t_k,s_k)\in\delta^* \\ k \neq i}} \theta_z(|i-k|)}. \tag{4.4}$$

The terms of the equation and their influence on the resulting selection are discussed below.

Distance function d

The function d represents the distance computed between two embeddings elements of different time periods. In our implementation, we use the Euclidean distance, for consistency with the distance metric employed in the K-means clustering of the MEI step.

d-weighting function β_z

The role of the β_z function is to control the proportion of scarce elements included in the optimal sequence, based on the parameter z. It takes as input the scarcity of an embedding element, and returns a z dependent weighting factor. An audio skim with high

faithfulness ($z = 0$) would heavily weight the distance between embedding elements with $s = 1$ within each time period, as these best represent the most present sound source of the full-length audio. In this case, the weights for all other elements should be 0. Conversely, an audio skim with high highlightness ($z = 1$) would assign equal weight to all elements, regardless of their scarcity score, ensuring maximum diversity in sequence selection. To support summaries with a balance between highlightness and faithfulness ($0 < z < 1$), the β_z function should also smoothly interpolate between these objectives. To reflect these requirements, β_z should be a decreasing function with respect to z and $s_{t,p}$, bijective with respect to z, that meets the following conditions:

$$\beta_z(s_{t,p}) = \begin{cases} 1 & \text{if } z = 1, \forall s_{t,p} \\ 1 & \text{if } z = 0 \text{ and } s_{t,p} = 1 \\ \approx 0 & \text{if } z = 0 \text{ and } s_{t,p} > 1 \end{cases} \tag{4.5}$$

Several function families meet these conditions, including rational decay, power-law decay, logarithmic decay, exponential decay, and composite functions. For this study, we select a composite function based on exponential decay:

$$\beta_z(s_{t,i}) = a_\beta^{-(1-z)^{k_\beta} \cdot (s_{t,i}-1)}, \quad \text{where } \begin{cases} a_\beta > 1, \\ k_\beta \geq 0 \end{cases} \tag{4.6}$$

This formulation is chosen because it conveniently provides controlled provides controlled decay for $z = 0$, ensuring that $\beta_0(s_{t,p} > 1) \geq 1/a_\beta$. By selecting an appropriate a_β, we can thus satisfy the condition $\beta_1(s_{t,p} > 1) \approx 0$. The parameter k_β adjusts the shape of the decay, allowing a smooth trade-off between faithfulness and highlightness through the parameter z. An illustration for $a_\theta = 100$, and $k_{beta} = 6$ is shown in figure 4.2.

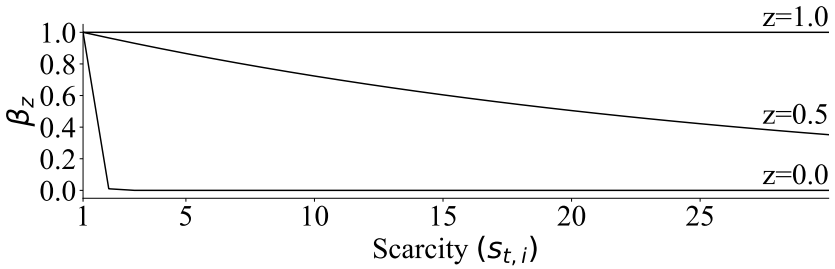


Figure 4.2 – $Beta_z$ weight for different z values, with $a_\beta = 100$, and $k_\beta = 7$.

d-weighting function θ_z

The role of θ_z is to influence how well the chosen sequence preserves the chronological structure of the full-length audio, by weighting the distance between pairwise elements according to their relative position within the sequence. This function takes as input the absolute difference between the positions of two elements in the sequence, and returns a z dependent weighting factor. An audio skim with high highlightness (z=1) would weight each pairwise distance evenly, as it prioritizes diversity of events over their temporal order. An audio skim with high faithfulness (z=0) would place maximum emphasis on first-order neighbors, and ignore contributions from more distant pairs. This maintains saliency while adaptively compressing time when $N < T$, without excluding the possibility of similar elements occurring in the sequence. As a result, θ_z should also be a decreasing function with respect to both z and $s_{t,p}$, bijective with respect to z, and needs to satisfy the same conditions as β_z presented in Equation 4.5. We thus adopt the same composite function defined in Equation 4.6 for θ_z, with parameters a_θ and k_θ.

Elements near the middle of the sequence have more higher-order neighbors than those positioned at the boundaries, leading to an edge effect where time periods at the beginning or end receive lower cumulative θ_z weights. To ensure a balanced weighting distribution across the audio skim, the ω_z formulation normalizes θ_z by dividing it by the sum of θ_z weights over all available neighbors in the sequence.

Complexity of exhaustive optimization of SME

We analyze the computational complexity of solving the optimization objective by exhaustively evaluating all possible combinations. There are $\binom{T}{N}$ possible combinations of N time periods selected from T. There are k_t element in each period t, and we define $k_{max} = \max_{t \in \{1,2,\ldots,T\}} k_t$. For each combination, a double summation is performed across its N elements, resulting in N^2 computations of the distance d. The overall complexity is thus $O(N^2 \binom{T}{N} * {k_{max}}^N)$.

For instance, with $T = 24$ time periods, $N = 12$ selected periods, and $k_{max} = 10$ samples per period:

— $\binom{T}{N} = \binom{24}{12} = 10^6$ combinations

— $k_{max}^N = 10^{12}$ combinations

This results in approximately 10^{18} combinations, leading to 10^{20} distance computations for a brute-force approach that evaluates every distance to find the optimal solution. With

each computation taking roughly 10^{-5} s on a CPU for embeddings of size 2048, the total runtime would exceed 10^{15} s, which represents more than a million years. We therefore adopt a greedy-based algorithm to approximate the optimal solution.

Greedy Approach

To address the complexity issue, we propose a greedy algorithm to find an approximate solution to δ_{opt}, referred to as $\widehat{\delta}_{\text{opt}}$. The algorithm is detailed in Algorithm 1.

For initialization, we choose the embedding element that is the closest to the β_z-weighted average of all available embedding elements, ensuring that the first selected element is the most representative of the collection of elements identified in MEI. The algorithm then iteratively expands δ_{opt}, by incrementaly adding the element that maximizes the ω-weighted sum of distances to the elements already included in the sequence.

This reduces the complexity of the problem to:

$$O\left(\sum_{i=1}^{N} k \cdot k_{\max} \cdot (T - i + 1)\right) = O(N \cdot k_{\max} \cdot T)$$

This algorithm makes the problem manageable, as in the example provided in Section 4.2.3, it results in approximately $12 \cdot 24 \cdot 10 \approx 10^3$ distance computations.

4.3 Experimental protocol

4.3.1 Evaluation Metrics

To allow the objective evaluation of audio skims, we define two quantitative metrics that separately measure faithfulness and highlightness. Since these two metrics are combined to compute the final performance metric, a calibration step ensures that they vary within a comparable range. To measure the controllability of the trade-off between faithfulness and highlightness, we also introduce the z-controllability metric.

Faithfulness objective Metric

To quantify faithfulness, we propose to consider the Dynamic Time Warping Inverse Distance (DTWID), computed between the L-CLAP embeddings of the full-length recording and the ones of the audio skim. For the audio skim, L-CLAP embeddings are computed on 8-second segments with 5-second overlap. For the full-length audio, L-CLAP embeddings are computed on non-overlapping 8-second segments.

Algorithm 1 Greedy algorithm for δ_{opt}

Input: N(length of sequence), z (user input), output of MEI ($e_{t,p}$ and $s_{t,p}$ for $t \in [1, T]$ and $p \in [1, k_t]$), d (distance function)
Step 1: Initialisation
Create Sequence: $\widehat{\delta}_{\text{opt}} \leftarrow \emptyset$
Compute the weighted average:

$$\mu = \frac{\sum_{t=1}^{T} \sum_{p=1}^{k_t} \beta_z(s_{t,p}) e_{t,p}}{\sum_{t=1}^{T} \sum_{p=1}^{k_t} \beta_z(s_{t,p})}.$$

Select the embedding $e_{t_1,p_1} \in \Gamma$ that is closest to μ:

$$e_{t_1,p_1} = \underset{e_{t,p}}{\operatorname{argmin}}\, d(e_{t,p}, \mu).$$

Add (t_1, p_1) to $\widehat{\delta}_{\text{opt}}$.
Step 2: Iterative selection using heuristic
while $|\widehat{\delta}_{\text{opt}}| < N$ **do**
Compute the next sample:

$$e_{t_n,p_n} = \underset{\substack{e_{t_i,p_i} \\ t_i \notin \delta_{\text{cur}}[1]}}{\operatorname{argmax}} \sum_{(t_j,p_j) \in \widehat{\delta}_{\text{opt}}} \omega_z(s_{t_i,p_i}, s_{t_j,p_j}) d(e_{t_i,p_i}, e_{t_j,p_j})$$

Add (t_n, p_n) to $\widehat{\delta}_{\text{opt}}$.
Sort $\widehat{\delta}_{\text{opt}}$ by t value.
end while
Step 3: Terminate
Return $\widehat{\delta}_{\text{opt}}$.

The dynamic time warping allows to take into account faithfulness, as it aligns two sequences by non-linearly stretching or compressing the time axis to measure their similarity. The DTW distance is inverted so that the final score ranges between 0 and 1, where 1 indicates an audio skim that most closely matches the background sounds and temporal structure of the full-length audio.

Direct computation of Dynamic Time Warping Inverse Distance (DTWID) between the L-CLAP embeddings of the full-length recording and the audio skim is not feasible, as a typical 1min audio skim would contain 12 audio segments of 8s with 3s overlap, while a 24-hour recording consists of approximately 10,000 audio segments of 8s. Since DTW is typically computed on sequences of comparable lengths, the full-length audio is splitted into segments within 15min time intervals, and the DTW is computed with the average of L-CLAP embeddings over each time period. This produces 96 embedding elements for the full-length audio and 12 elements for the skim, making DTWID computation tractable.

A reference audio skim with a focus on faithfulness might be constructed by dividing the full-length audio into several time intervals, selecting the segment containing the most present sound in each interval, and concatenating these segments in chronological order. We use this principle to generate a calibration audio skim, against which the faithfulness of other skims is measured. This audio skim is constructed by splitting the full-length audio into N time periods, computing the mean of L-CLAP embeddings over those time periods, and selecting N embedding elements over all the full-length audio that are the closest to those mean values. The faithfulness objective metric is then defined as the ratio of the DTWID score of the evaluated audio skim to that of the calibration audio skim:

$$\text{Objective Faithfulness}(x) = \frac{\text{DTWID}(x)}{\text{DTWID(reference)}} \tag{4.7}$$

Highlightness objective Metric

To quantify highlightness, we compute the Intersection over Union (IoU) of class activations from a multi-class audio classifier. This measures the overlap between detected sound classes in the audio skim and the full-length recording across different confidence thresholds. This similarity metric, based on the maximum logit values, helps identify the most prominent sound sources in both the audio skim and the full-length audio and assesses whether they are consistent between the two. Specifically, we employ BEATs (S. Chen et al. 2022), an attention-based audio classifier trained on Audioset (Gemmeke et al. 2017). BEATs is a general-purpose classifier with 527 output classes from the Audioset

ontology (Gemmeke et al. 2017), covering a wide range of sound sources. BEATs logits are computed on 8-second segments with a 5-second overlap in the audio skim, while for the full-length audio, they are computed on non-overlapping 8-second segments. Let p_i denote the maximum output of the BEATs classifier for class i across the audio skim, and q_i denote the maximum output for class i across the full-length audio. Define $1_{[th,1]}(x)$ as the indicator function, which equals 1 if $x \in [th, 1]$ and 0 otherwise. The CIoU (Classification Intersection over Union) is then defined as:

$$\text{CIoU(th)} = \frac{\sum_{i=1}^{N} 1_{[th,1]}(p_i) \cdot 1_{[th,1]}(q_i)}{\sum_{i=1}^{N} \left(1_{[th,1]}(p_i) + 1_{[th,1]}(q_i) - 1_{[th,1]}(p_i) \cdot 1_{[th,1]}(q_i)\right)} \tag{4.8}$$

To account for diverse audio environments, where some are very calm and others have prominent sound sources that may result in high output logits, we compute the Area Under the Curve of the CIoU for varying thresholds (CIoUAUC):

$$\text{CIoU-AUC} = \int_{0^+}^{1^-} \text{CIoU}(x)\, dx \tag{4.9}$$

A reference audio skim oriented toward highlightness could be obtained by making a selection of segments that capture the most diverse and salient sound sources, regardless of their position in time. We use this principle to generate a calibration audio skim, against which the highlightness of other skims is measured. This audio skim is produced by selecting the top N audio samples that maximize the top N BEATs logit outputs across the full-length audio. Audio samples are selected incrementally, from the ones corresponding to the highest-ranking BEAT logit, to the one corresponding to the lowest within the top N. If an audio sample maximizes k different BEAT logits among the top N, the top N BEATs list is extended to a top N+k for the next samples. The CIoUAUC is then computed for this calibration summary. The objective highlightness metric is defined as the ratio of the CIoUAUC of the evaluated summary to that of the calibration summary:

$$\text{Objective Highlightness}(x) = \frac{\text{CIoU-AUC}(x)}{\text{CIoU-AUC(reference)}} \tag{4.10}$$

In practice, we discretize the CIoUAUC integral of Equation 4.10 by evaluating CIoU across thresholds ranging from 0 to 1 in steps of 0.01. Instances where the union is 0 are excluded from the sum to avoid division by zero in Equation 4.8.

Performance metric

Let us consider a 2d-plane with faithfulness and highlightness as X and Y axes. We define the **performance** of an audio skim generation algorithm as the Area Under the Curve (AUC) on this plane across a sequence of z values. If only one value is available, meaning that the method can't produce multiple audio skims for different z values, the AUC is defined as the area of the rectangle formed by the axes and the single point. When multiple values are available, we reorder points by increasing faithfulness and connect them with straight-line segments. A horizontal line is added from the highlightness axis to the first point, and a vertical line from the last point to the faithfulness axis. The AUC is then computed under this resulting curve. An example is shown in Figure 4.3.

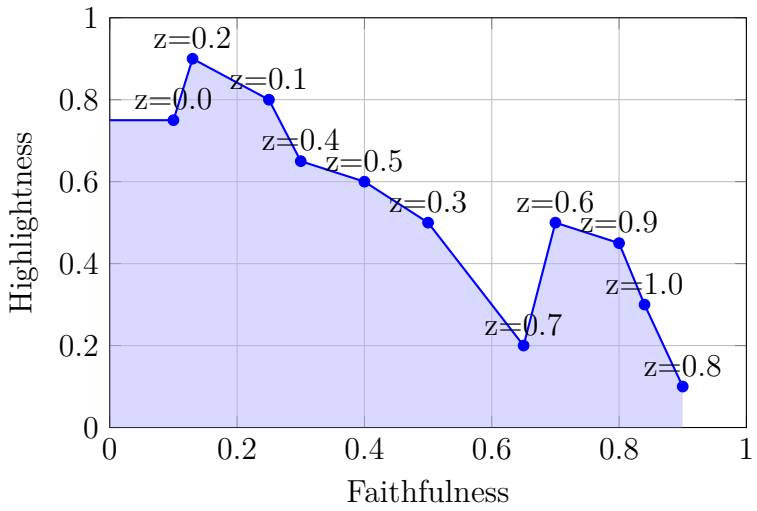


Figure 4.3 – Illustration of the performance metric in the faithfulness/highlightness plane with different z values. The area under the blue curve represents the performance of the audio skim generation algorithm.

z-controllability

While the performance metric evaluates the overall quality of an audio skim generation algorithm, it does not indicate how effectively the algorithm controls the trade-off between faithfulness and highlightness using z as an input. To quantify this, we introduce the z-controllability ρ_z, defined as:

$$\rho_z = \frac{1}{2}\left(\rho(z, highlightness) - \rho(z, faithfulness)\right)$$

where $\rho(z, highlightness)$ is the Pearson correlation coefficient between z and a highlightness score, and $\rho(z, faithfulness)$ is the Pearson correlation coefficient between z

and a faithfulness score. The z-controllability metric can only be computed if the model generates distinct summaries for varying values of z.

4.3.2 Dataset

We conduct 24-hour continuous recordings (from midnight to midnight, local time) in four distinct urban environments: city center, city neighborhood, residential area, and peri-urban area. For each environment, we find two suitable locations, resulting in a total of eight recording sites. At each location, we deploy an AudioMoth sensor (Hill et al. 2018), positioning it on a window, balcony, or garden. We record audio in 16-bit, 32 kHz mono. Each location is recorded for two full days, from midnight to midnight, which are not necessarily consecutive.

For our evaluation set, we select recordings from 4 different locations (one per environment). For each location, one of the 2 available recordings is used to manually produce reference audio skims (created by the author, a trained sound engineer) and to conduct a subjective evaluation. For model validation, we use recordings from the remaining four locations. A detailed listing of all recordings is provided in Table 4.1.

Environment Type	**Location ID**	**Date**	**Split**
City Center	1	2025-01-06	eval.+subj.eval.
		2025-01-07	eval.
	2	2024-12-31	valid.
		2025-01-01	valid.
City Neighborhood	3	2024-12-09	eval.+subj.eval.
		2025-01-03	eval.
	4	2024-12-17	valid.
		2025-01-08	valid.
Residential Area	5	2024-12-27	eval.+subj.eval.
		2024-12-28	eval.
	6	2025-01-13	valid.
		2025-01-14	valid.
Peri-Urban	7	2024-12-07	eval.+subj.eval.
		2024-12-08	eval.
	8	2024-12-22	valid.
		2024-12-23	valid.

Table 4.1 – Overview of recordings with environment types, locations, dates, and dataset split.

The 8 24-hour recordings in the evaluation and validation set are insufficient to support strong statistically meaningful conclusions. To expand the dataset and strengthen the evaluation using our objective metrics, we sample 12h full-length audio recordings from the 24h recordings to supplement our dataset. We apply a 3h overlap when segmenting the 24h recordings, resulting in 5 12h recordings per 24h recording. Consequently, for the 16 24h recordings in the evaluation and validation set, this procedure yields 80 different 12h recordings.

4.3.3 Reference Audio Skims

The author manually produced audio skims for the 4 24h full-length audio labeled as *subj.eval.* in Table 4.1. For each full-length audio, three skims were created, corresponding to different values of the user-defined trade-off parameter z: z=0, z=0.5, and z=1. The sound engineer browsed through the full-length audio using a digital audio workstation (DAW) with both waveform and spectrogram views. This audio skim generation process takes approximately 7h per full-length audio.

4.3.4 Baselines

We compare our model's results to two baseline approaches:

— **Random**: selects a random sequence of $N = 12$ audio segments (8s each) from the full-length recording, not necessarily in chronological order.

— **Downsampling**: The full-length audio is divided into equal time periods (2 hours for 24h recordings and 1 hour for 12h recordings). An 8s audio segment is then randomly selected within each period, for a total of $N = 12$ segments. This random selection helps avoid the artificial inclusion of periodic events such as church bells, which ring every hour and would otherwise appear in every sample between 9a.m. and 7p.m.

4.3.5 Experiment settings

As explained in Section 4.2.2, the texture window of each audio segment used for embedding computation in the MEI step is fixed to 8s. For the summaries, we set $N = 12$ with a 3s overlap between consecutive elements, resulting in a total skim duration of

1min3s. The full-length recordings are divided into 15min time periods, corresponding to $T = 96$.

MEI settings

Clustering is performed using the K-means algorithm on L-CLAP embeddings, with a maximum of 300 iterations and K-means++ initialization (Arthur et al. 2006).

MEI and SME parameters tuning

In Table 4.2, we present the correspondence between the high-level task description and the parameters and weights used in the automatic audio skim generation pipeline. We fine-tune our parameters using the 40 overlapping 12h full-length audios of validation presented in section 4.3.2. We divide the parameters into three groups: the first is optimized for performance, the second for contrast between $z = 0$ and $z = 1$, and the second for z-controllability, as defined in Section 4.3.1. Accordingly, we adopt a three-stage optimization strategy where sets of parameters are optimized sequentially. We first optimize for performance, as the quality of the summaries is our primary concern. For parameters yielding equivalent performance, we then select those that maximize the contrast between the two extremes of z. Finally, within this range, we choose parameters that provide users with the most linear control possible over the trade-off between highlightness and faithfulness.

- Optimization of parameters for performance (CR, a_θ, and a_β) .

In this optimization step, we evaluate performance as defined in Section 4.3.1, based on the objective faithfulness and highlightness metrics. The parameters CR, a_θ and $a\beta$ have a direct influence on the performance metric. Since performance is our primary evaluation criterion, we begin by fine-tuning these parameters to maximize it, focusing on the two endpoints of the trade-off curve, z=0 and z=1. A grid search over CR, a_β and a_θ, reveals that maximum performance occurs within the range CR=5 to CR=95, with no significant variation accross this interval. There is no significant improvement in the performance metric for $a_\theta \geq 5$ and $a_\beta \geq 5$.

- Optimization of parameters for trade-off contrast (CR, a_θ, and a_β) .

To select an optimal value, we choose the minimum CR that maximizes the distance between the two endpoints on the faithfulness/highlightness plane, as a greater distance

Step	Operation	Description	Param.	perform.	contr.	z-control.	selection
MEI	T-period clustering	Split the full-length audio into T time periods, and select $CR\%$ of the elements from each time period	CR	$\geq 5, \leq 95$	15	–	15
SME	β_z weighting	Modulates the number of scarce elements that will be found in the final summary, depending on z	a_β	≥ 5	≥ 5	5	5
			k_β	–	6	6	6
	θ_z weighting	Modulates the diversity of the sequence based on how far apart pairwise elements are in the sequence, depending on z	a_θ	≥ 5	≥ 5	10	10
			k_θ	–	3	3	3

Table 4.2 – MEI and SME parameter description, tuning outcomes, and selected values

enables the generation of more contrasted summaries. The optimal parameter is found to be CR=15. There is no significant improvement of this contrast for $a_\theta \geq 5$ and $a_\beta \geq 5$.

- Optimization of parameters for z-controllability (a_θ, a_β, k_θ, k_β) .

We next fine-tune the parameters that control the shape of the β_z and θ_z weight function, namely a_θ, a_β, k_θ, k_β. These parameters control how the summary transitions between high highlightness and high faithfulness across the full range of the trade-off parameter z, from 0 to 1. Since these parameters are designed to influence how the summary evolves on the faithfulness/highlightness plane between the endpoints of z ($z = 0$ and $z = 1$), we tune them based solely on the z-controllability criteria. We determine that the highest z-controllability score is achieved with a_β=5, a_θ=10, k_β=6, k_θ=3.

4.3.6 Computational efficiency of the pipeline

On a GeForce RTX 2080 Ti GPU, the MEI stage takes 420s to compute embeddings for a 24h full-length audio and 7s for per-period clustering. The SME stage then takes an additional 16s on CPU. The MEI stage needs to be run only once per full-length audio. Generating multiple summaries for different z values thereafter requires only the 16s SME step.

4.3.7 Evaluation protocol

Objective evaluation

As explained in Section 4.3.2, the objective evaluation is conducted on the 40 12-hour recordings of the evaluation subset, denoted as *eval.* in Table 4.1. The objective faithfulness and highlightness metrics introduced in Section 4.3.1 are used to compute the objective performance metric. In addition, for the random and downsampling baselines, we repeat the evaluation with 20 different random seeds to account for variability. For the SkimEAR method, summaries are generated with z values ranging from 0 to 1 in increments of 0.1.

Subjective evaluation

We have recruited 4 experts (acoustic, and signal processing researchers) to evaluate faithfulness and highlightness of audio skims. Each expert assesses summaries from 4 full-length recordings, those for which expert reference skims are available and labeled as *subj.eval.* in Table 4.1. They evaluate a total of 8 different skims per full-length audio:

— The three expert reference skims produced by a sound engineer, as presented in Section 4.3.3,

— The two baseline skims, as described in Section 4.3.4,

— The three automatic skims generated using our method described in Section 4.2, with z=0, z=0.5 and z=1.

Experts are provided with definitions of faithfulness and highlightness, along with example audio skims and their expected assessments. They conduct the evaluation using Audacity software [2] with a spectrogram view to explore both the full-length recordings

2. Software available at: https://audacity.fr/

and their corresponding skims. Ratings are performed by placing a cross on a 2D plane, where the x-axis represents faithfulness and the y-axis represents highlightness. Experts are given seven skims produced from a full-length recording in the validation set. These skims are generated by the author of this manuscript, using the same method described in Section 4.3.3.

4.4 Results

4.4.1 Objective evaluation

Table 4.3 reports the averaged outcomes across all recordings, including averages over seeds for the baseline conditions. The relative positions of the different methods across z values on the highlightness/faithfulness plane are illustrated in Figure 4.4.

Both the random and downsampling baselines yield poor performance scores, as the performance metric is computed from a single value. Downsampling exhibits greater faithfulness than pure random selection, likely due to the fact that the selected samples, although random within each time period, still follow the chronological order of the original recording. Interestingly, these baselines tend to lean more towards faithfulness. This can be explained by the fact that randomly selected segments are statistically more likely to reflect the most present sound sources within the full-length audio. Among all methods, the downsampling method results in the highest faithfulness score overall. This outcome is consistent with the fact that downsampling is closest to the experience sampling method discussed in Section 4.1, which served as a basis for defining faithfulness.

On the 40 12h recordings, our method achieves a z-correlation score of 0.78, indicating strong controllability over the highlightness/faithfulness trade-off. For the baseline methods, z-controllability can't be computed, as they provide summaries for only one z value, which doesn't allow a Pearson-r correlation computation.

Method	**Objective Performance** *Highlightness/Faithfulness AUC* ↑
Random	.15 (±.15)
Downsampling	.18 (±.17)
SkimEAR	**.34** (±.17)

Table 4.3 – Comparison of audio skim generation methods for the 40 12-hour recordings of the evaluation set, with 95% confidence interval. Bold values indicate results not significantly different from the best-performing condition, according to paired t-test ($p < 0.05$).

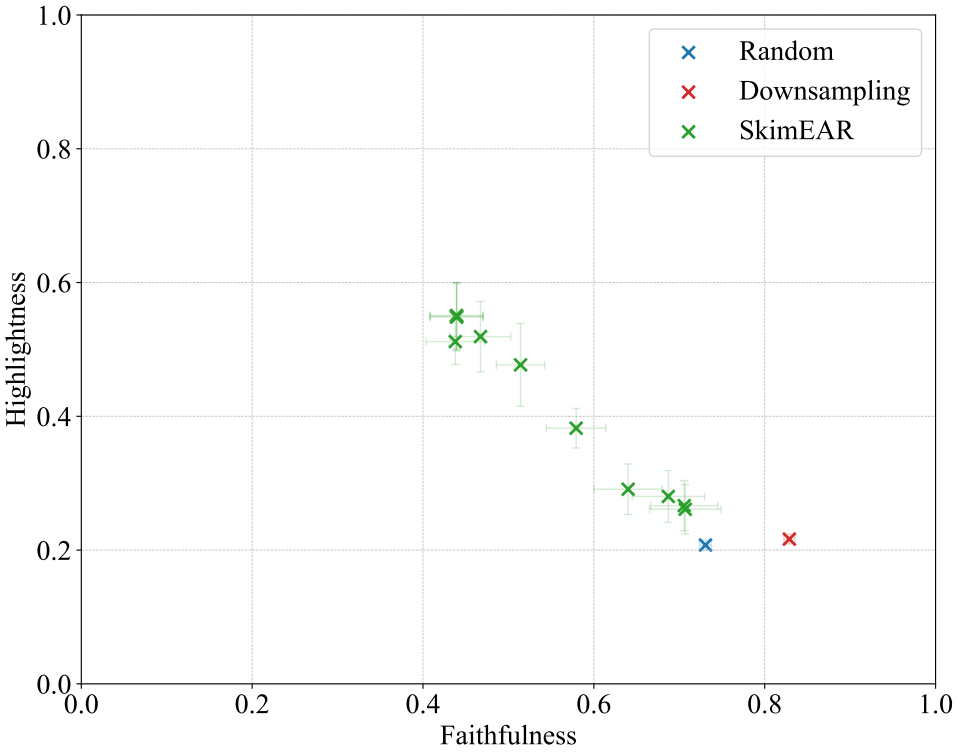


Figure 4.4 – Objective faithfulness and highlightness for audio skim generation methods. Each point represents an average over the evaluation datasets, and error bars indicate 95% confidence intervals. Results are averaged over the 40 12h recordings in the evaluation set. For the *Random* and *Downsample* methods, results are also averaged over 20 different seeds. For *SkimEAR* method, each point corresponds to a different value of the z parameter.

4.4.2 Subjective evaluation

Table 4.3 reports the averaged outcomes across all recordings used for subjective evaluation. For comparison, we also computed the objective performance scores using the faithfulness and highlightness metrics on the same subset of audios. These scores are of the same order of magnitude as those reported in the previous section, suggesting that the subset of recordings used for the subjective evaluation is representative of the dataset employed in the objective evaluation.

SkimEAR method achieves subjective performance scores not significantly different from those of the expert reference skims (with a paired t-test and p-value threshold of 0.05), while remaining significantly better than the baseline methods. For both the expert reference and our method, the relative positions of the z values on the faithfulness–highlightness plane follows the expected trend: $z = 0$ yields high faithfulness, $z = 1$ yields high highlightness, and $z = 0.5$ lies between the two. The expert reference skims span a broader range along both dimensions, achieving higher faithfulness at $z = 0$ and higher highlightness at $z = 1$ than our method, indicating that further improvement is possible.

z-controllability can't be computed for the subjective evaluation, as this evaluation is performed on a maximum of 3 different z value per method, which is insufficient to allow a meaningful Pearson-r correlation computation.

Method	**Objective Performance** *Highlightness/Faithfulness AUC* ↑	**Subjective Performance** *Highlightness/Faithfulness AUC* ↑
Random	.14 (±.16)	.13 (±.09)
Downsampling	.18 (±.17)	.16 (±.06)
SkimEAR	**.32** (±.05)	**.38** (±.16)
Expert reference	**.35** (±.13)	**.49** (±.08)

Table 4.4 – Comparison of the objective and subjective performances of audio skim generation methods for the 4 24-hour recordings of the subjective evaluation set, with 95% confidence interval. Bold values indicate results not significantly different from the best-performing condition, according to paired t-test ($p < 0.05$).

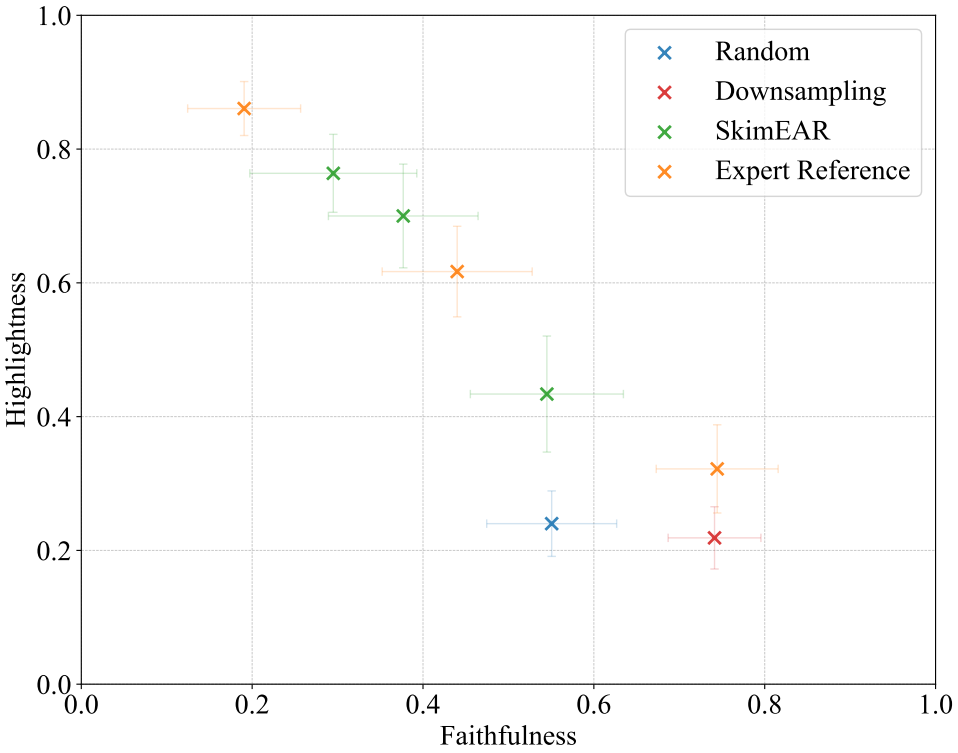


Figure 4.5 – Subjective faithfulness and highlightness for all audio skim generation methods. Each point represents an average over the evaluation datasets, and error bars indicate 95% confidence intervals. Results are averaged over the 4 24-hour recordings in the subjective evaluation set. For the *Expert reference* and *SkimEAR* methods, each point corresponds to a different value of the z parameter.

4.5 Discussions

4.5.1 Regarding the relevance of the objective metrics

While our two proposed metrics are, by design, intuitively aligned with the concepts of faithfulness and highlightness, their ability to reflect human perception of these notions has not yet been formally validated. We leverage the rater evaluations collected in Section 4.3.1 to assess how well our objective metrics correlate with perceived faithfulness and highlightness. We observe an average per-dataset, per-rater correlation of of 0.46 ± 0.12 for faithfulness and 0.58 ± 0.07 for highlightness (n= 8), with 95% confidence interval. The lower correlation observed for the faithfulness metric may be due to the fact that faithfulness is a more complex and less intuitive concept than highlightness, potentially leading to greater variability or misinterpretation among raters. Nevertheless,

these results indicate that both metrics are meaningfully aligned with human perception, supporting their relevance for an objective evaluation of audio skim generation. In future work, the perceptual ratings collected in this study could serve as a basis for designing new objective metrics that are even more closely aligned with human judgments.

4.5.2 Privacy considerations

Our subjective evaluation involves both the creation of expert reference summaries and the review of generated audio skims by domain experts. To conduct this evaluation meaningfully, experts must be able to listen to the corresponding full-length recordings and develop a clear understanding of the various acoustic scenes they contain. Speech plays a particularly important role in these scenes, as the recordings are captured in urban environments where spoken voices can constitute a meaningful and sometimes prominent part of the soundscape. However, some of these recordings include clearly intelligible speech, which raises privacy concerns.

To address this issue, we applied on all the full-length audios a speech content privacy enforcement algorithm designed to make speech unintelligible while preserving both the integrity of the acoustic scene, and the overall audio quality. We used the method described in (Tailleur, Lagrange, et al. 2025b), which first detects speech regions using Voice Activity Detection (VAD), then applies speech separation to isolate speech from background sounds. The isolated speech signal is segmented using low root-mean-square energy regions as boundaries, and the waveform within each segment is reversed, producing a time-inverted speech signal. Evaluated on a dataset that contains mixtures of speech and environmental background sounds, this algorithm has proven to successfully enforce speech content privacy while preserving both sound source detectability and audio quality.

This simple strategy is not intended to be fully robust, as the data are assumed to remain within a small group of experts. Under targeted attack scenarios, partial recovery of speech content may still be possible. Moreover, the method alters only the speech content while preserving speaker timbre, enabling the differentiation of emotions, gender, and age during listening. As a result, speaker re-identification may still be possible. For these reasons, the audio dataset used for the audio skim generation is not publicly released. Annotators are also instructed to delete all audio files from their local storage after completing the annotation process.

4.5.3 Considering more complex algorithms for addressing the SME problem

To improve the SME algorithm, we explored modifications to the element selection strategy. Rather than adding one element at a time, we experimented with selecting batches of two elements per iteration. However, this approach did not lead to any significant improvement in the performance metric. Given its higher computational cost, we retained the simpler single-element selection strategy.

We also investigated a variant that combines elements removal and replacement. Starting from an audio skim built using the standard greedy algorithm, the method consists in iteratively removing the element whose exclusion least contributes to the total distance defined in Equation 4.3. After each removal, the greedy algorithm is re-applied to potentially insert a better element. This process is repeated until no further improvement can be found. Like the batch selection method, this strategy did not improve the performance metric and, as it increases inference time, was not included in the final version of the algorithm.

These results suggest that simple variations of the greedy strategy, whether through batch selection or iterative refinement, are unlikely to yield substantial improvements. To move beyond these limitations, a fundamentally different approach may be required, potentially one that formulates the audio skim generation problem as a single-stage optimization rather than relying on iterative greedy selection.

4.5.4 Interactive refinement of the audio skim

The greedy algorithm constructs the audio skim incrementally, making it suitable for semi-automatic and user-guided generation. For example users could generate an initial audio skim automatically and then make adjustments, such as rejecting specific audio segments they find unsatisfactory. The algorithm can then propose alternative segments to replace the rejected ones. Users could also adjust the trade-off parameter z between the initial and refined versions, depending on whether they want the replacements to emphasize faithfulness or highlightness.

4.5.5 Audio skims from discontinuous recordings

For evaluation consistency, the audio skims in this chapter are produced from continuous 24h recordings, spanning midnight to midnight. However, the MEI and SME stages can equally be applied to discontinuous recordings, i.e., audio captured intermittently. In such cases, the Clustering Ratio (CR) naturally leads to a smaller number of clusters k_t for time periods t containing less audio. The SME step remains applicable under these conditions and continues to produce meaningful results.

4.5.6 Cyclic audio skims

The algorithm presented in Section 4.2 also supports an additional constraint to produce cyclic audio skims. We define a cyclic audio skim as one that is intended to be listened to in a continuous loop. This is particularly relevant for full-length recordings with inherent periodicity, such as 24-hour, weekly, or yearly recordings, where the beginning and end are expected to connect seamlessly.

To account for this cyclic structure, Equation 4.3 can be slightly adapted. The θ_z function in Equation 4.4 can be computed with modified inputs. In particular, the distance between the indices of the last and first elements should be small to reflect the cyclic nature of the full-length audio. The distance $|i-j|$ can thus be replaced by $\min(|i-j|, N-|i-j|)$. This modification ensures that neighborhood relationships between the elements of a sequence are treated cyclically rather than linearly.

4.5.7 Audio skims for temporal profiles

In this study, we have treated audio skims as shortened versions of full-length recordings. However, one could also imagine skims designed not to shorten the recording, but to capture the profile of a specific temporal pattern. For instance, as explored in Chapter 3 for source prediction, recordings may span several days or even months, and it could be useful to generate an audio skim representing an average day or average week. The SkimEAR pipeline can easily be adapted to produce such skims by modifying the MEI step. To construct a daily profile, for example, we would keep the same time intervals as in the standard MEI step, but aggregate data from all days of the recording for each interval. Thus, for a given time interval (e.g. between 7:15 and 7:30 a.m.), all corresponding audio segments from every available day would be grouped together, and K-means clustering would be applied to the embeddings of this set. Whether such profile-based

skims accurately represent the intended temporal patterns remains an open question for future evaluation.

4.5.8 Helping users interpret summaries

Since the generated summaries are not linear in time, they can at times be disorienting, as listeners may lose track of where they are temporally within the original recording. To address this, we propose to complement the audio skim with a visualization that indicates the corresponding time segments. [3]

4.6 Conclusion

In this work, we introduced a framework for generating audio skims that balance two competing criteria: faithfulness, which reflects how closely the summary preserves the temporal structure of the full-length audio, and highlightness, which reflects how well the audio skim represents the most salient sound sources of the full-length audio. We proposed a two stage approach based on identification of main events, and their sequencing. The identification of main events applies K-means clustering on different time periods. For the sequencing, we propose an optimization objective, that allows the user to control wether he wants an audio skim that is leaning more towards a high faithfulness score, or a high highlightness score. Our proposed greedy algorithm is computationally efficient, supports interactive refinement, and can incorporate constraints such as cyclic continuity.

Results on our evaluation dataset that contains 24h recordings from 4 different locations indicate that our method is performing similarly than summaries produced manually by a sound engineer.

Future work could extend the method to even longer recordings and assess its applicability to FTO data commonly used in urban acoustic monitoring. Although designed for audio, we believe that the method is quite general and can be adapted to any form of sequential data, including video skim generation tasks.

3. Examples of visualizations designed to accompany audio skims are available at: https://modantailleur.github.io/paperAudioSummary/

Takeaways and Perspectives

This chapter provides an answer to our third research question:

RQ3

How to provide meaningful audio skims of long audio recordings?

Key Takeaways

— We proposed a novel two-stage framework for generating audio skims that balance faithfulness and highlightness, two competing objectives reflecting background sound and temporal structure preservation, and sound source saliency and diversity representativeness.

— The framework comprises 2 stages: 1) identification of main audio events through K-means clustering on segmented time periods and 2) sequencing of these events via an optimization objective that enables user-controlled trade-offs between faithfulness and highlightness.

— Our greedy algorithm used for sequencing is computationally efficient, supports interactive refinement, and can incorporate practical constraints such as cyclic continuity in audio skims.

— Evaluation on 24-hour recordings from multiple locations shows that our method performs comparably to manually produced summaries by a sound engineer.

In the near future, and building on the results presented in Chapter 3, audio skims could be integrated into sound visualization pipelines, enriching the user experience and providing a novel way to explore the soundscape. Interviews conducted in (Tailleur, Aumond, Tourre, et al. 2024) suggest that this approach is particularly relevant for communicating about soundscapes with the general public.

Chapter 5

Discussion

5.1 PANN-FTO: new applications and perspectives for improvement

5.1.1 Using FTOs and sound sources predictions for other applications than urban environmental sound recordings

This section discusses outcomes of the research described in the main-authored article "Machine listening in a neonatal intensive care unit", published at DCASE 2024 (Tailleur, Lostanlen, et al. 2024).

As PANN-FTO produces predictions for over 500 sound source classes, its potential thus extends beyond urban soundscape analysis. In particular, in (ibid.), we deployed sensors capable of recording FTOs in a neonatal intensive care unit (NICU). In this context, FTOs serve primarily as a privacy-preserving alternative, since sensors are deployed in semi-private rooms shared by the infant and their family during care. Interviews with parents revealed that the NICU environment is perceived as particularly noisy, and that it influences how parents care for their infants. We identified four types of sound events that characterize the acoustic environment in NICUs: adult voices, footsteps, oxygenators, and alarm devices.

A key challenge lies in the specificity of these sounds, which may not align directly with the AudioSet ontology used by PANN-FTO. To explore possible mappings, we ran the PANN-FTO model on FTO data from the NICU and looked for some frequently occuring AudioSet classes. We found four activity patterns of interest: "conversation", "walk, footsteps", "train", and "electronic music". Although the former two sound events are plausible, the latter two are clearly not. Yet, after interviewing NICU employees, we may hypothesize that they yield indirect information: i.e., that "train" actually corresponds to

Neonatal Intensive Care Unit (NICU)	AudioSet
Conversation	Conversation
Footsteps	Walk, footsteps
Oxygenator	Train
Hospital phone	Electronic music

Table 5.1 – Mapping of sound event labels from the neonatal intensive care unit (NICU) to AudioSet.

the rumble of the oxygenator while "electronic music" corresponds to the ringtone of the hospital phone. We summarize this correspondence in Table 5.1.

As in the Cense Lorient dataset, we faced the constraint that FTOs can't be listened to directly. To evaluate prediction reliability, we instead leveraged badge data from healthcare professionals and mothers, which logged their presence in the room. We found that detected sound events aligned well with these presence patterns: footsteps were typically detected when a single person was present, while conversations were detected when two or more individuals were in the room. These findings suggest that FTO recordings, combined with the PANN-FTO model, can be useful in a variety of contexts beyond urban environments, such as bioacoustics or industrial monitoring, particularly if privacy and efficient data transfer are required.

5.1.2 Revisiting classifier choice for transcoder-based sound source inference

Recent audio classifiers such as BEATs (S. Chen et al. 2022), which was used in Chapter 4, could theoretically be integrated into the transcoder architecture described in Section 2.2, replacing PANNs. BEATs has been shown to consistently outperform most other open-source classifiers on general-purpose benchmarks such as AudioSet (Gemmeke et al. 2017) and ESC-50 (Piczak 2015). BEATs expects a different input representation than PANNs, with higher spectral resolution (128 Mel bands versus 64 for PANNs), which is incompatible with the current transcoder design. Integrating BEATs would therefore require re-training a new transcoder capable of producing higher-resolution spectrograms. Although this task is relatively fast and straightforward, as explained in Chapter 2, it becomes a $\times 54$ super-resolution problem in the Mel domain, compared to the $\times 27$ super-resolution used for PANNs. As a result, it remains uncertain whether this approach would

yield satisfactory performance.

AudioSet and ESC-50 benchmarks include a broad range of sound categories, such as musical instruments, that are not necessarily relevant to urban environmental sound monitoring. If such classifiers are to be considered for use in the transcoder, preliminary studies are necessary to evaluate whether they provide a significant improvement over PANNs for the specific sound sources of interest. For urban environmental sounds, such assessments could be conducted using datasets like SONYC-UST (Bello et al. 2019) or UrbanSound8K (Salamon et al. 2014), which are also used for the evaluation of the current transcoder in Section 2.2.

5.2 Reconstructing audio waveforms from FTOs

This section discusses outcomes of the research described in the main-authored article "Diffusion-based Spectral Super-Resolution of Third Octave Acoustic Sensor Data: Is Privacy at Risk ?" published at EUSIPCO 2025 (Tailleur, Benaatia, et al. 2025).

Predictions from PANN-FTO can sometimes be ambiguous, especially when the model activates unexpected sound classes that are clearly out of context, such as "waves" or "horse" in the middle of a dense urban area. This behavior likely results from the transcoder being trained exclusively on urban sounds, without exposure to the full diversity of acoustic scenes. As predictions at the frame level (every 10 seconds) tend to be noisy, it is often more meaningful to analyze general trends with PANN-FTO rather than individual activations. For example, the daily profiles shown in Chapters 2 and 3 provide reliable insights, as individual misclassifications tend to average out over time. However, in monitoring scenarios where a specific day or time period needs to be analyzed, frame-level predictions may still be needed. In these cases, ambiguous class activations are difficult to interpret, as it would be unclear whether PANN-FTO has made a misclassification or whether an unusual acoustic event has genuinely occurred, since FTOs are not directly listenable. To address this limitation, we propose in Section 5.2.1 a deep learning approach to reconstruct audio waveforms directly from FTOs. In Section 5.2.2, we show how this method can be used to generate audio skims from FTOs. Given the promising results, we revisit the risks of recovering speech content from FTOs in Section 5.2.3.



Figure 5.1 – Pipeline for audio super-resolution. Module outlined with dashes does not require any training, module with a plain outline is only pre-trained, and module outlined in bold is specifically trained for the task.

5.2.1 Diffusion-based spectrogram super-resolution for waveform reconstruction from FTOs

Previous work by Gontier et al. Gontier (2020) investigated waveform reconstruction from FTOs using deep learning. In their work, two deep learning approaches are proposed to reconstruct fine-band short-time fourier transform magnitudes, which are then converted into audio using iterative phase recovery. The first method refines a PINV spectral estimation by learning a corrective term, while the second introduces a multi-band architecture that accounts for frequency-dependent spectral decay and uses a Generative Adversarial Network (GAN) to produce more realistic high-frequency content. However, their reconstructions achieved only limited fidelity, likely due to training constraints.

Diffusion models have emerged as a robust alternative to GANs for generative tasks (Ho et al. 2020), offering high reconstruction quality in many applications. These models are easier to train than GANs and can thus be applied to a broader range of fields. They have demonstrated super-resolution abilities (Kawar et al. 2022; Saharia et al. 2022), and have shown promising results in audio enhancement tasks (Sawata et al. 2023; Serrà et al. 2022).

We therefore introduce a diffusion-based approach to transform FTOs into audio waveforms. Our proposed pipeline transform FTOs into Mel spectrograms using a super-resolution algorithm, and uses a vocoder for phase reconstruction. We refer to the super-resolution algorithm as Diffspec. By relying on a pre-trained vocoder for the vocoding stage, the model only needs to focus on spectrogram reconstruction, thereby simplifying the overall process. The pipeline is summarized in Figure 5.1.

To implement the Diffspec model, we adopted the super-resolution architecture proposed by Saharia et al. (Saharia et al. 2022), which we adapt for spectrogram reconstruction. The proposed method has notably been used in the NU-Wave 2 algorithm (Han et al. 2022). For the vocoder, we select the pre-trained GomiGAN model (M. Kang et al. 2023).

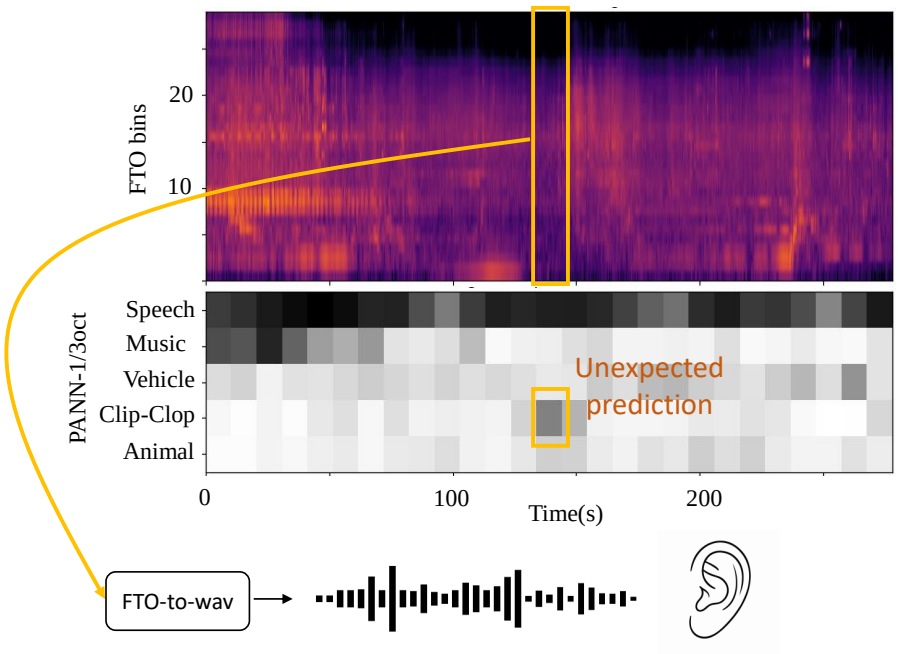


Figure 5.2 – Pipeline for sound analysis, using both PANN-FTO and Diffspec.

We train our model on TAU Urban Acoustic Scenes 2020 Mobile dataset (Mesaros, Heittola, and Virtanen 2018). This dataset contains 10s audio clips from 10 different acoustic scenes, including indoor public spaces, public transports, streets, and parks. It totals 64 hours of audio.

We refer to this full pipeline trained on TAU dataset as FTO-to-wav in the remainder of this manuscript. Informal listening suggests that the reconstructed audio is of high quality compared to using a simple PINV transcoder [1]. In future work, we aim to formally assess the reconstructions of FTO-to-wav. One possible evaluation strategy would be to apply PANNs models to the reconstructed waveforms and compare classification accuracy against that of PANN-FTO on the same datasets as the ones presented in Section 2.2.

FTO-to-wav is computationally expensive, requiring approximately 80 times more processing time than PANN-FTO inference. A practical compromise in using FTO-to-wav for monitoring is to use PANN-FTO for a first-pass analysis, and then apply the slower FTO-to-wav reconstruction selectively, only in time segments where PANN-FTO class activations appear ambiguous. This approach is illustrated in Figure 5.2.

1. Audio examples available at: `https://modantailleur.github.io/fast-to-wav/`

5.2.2 Audio skim generation from FTOs

In Chapter 4, we proposed an audio skim generation pipeline that allows the transformation of a 24h audio recording into a 1min summary. The core of that pipeline is an embedding generation used for K-means clustering and for segment selection. When full-band audio is available, the embeddings are computed directly from the waveform. However, as explained in Chapter 1, there are valid motivations for recording FTOs instead of raw audio waveform for environmental acoustic monitoring, mainly concerning frugality and privacy preservation. If only FTO measurements are stored, PANN-FTO outputs can serve as a source of embeddings. In other words, the pipeline shown in Figure 4.1 remains unchanged when using FTOs: it simply produces a 1min skim in the FTO domain, which can then be rendered as an audible summary via the FTO-to-wav reconstruction.

Evaluating this approach, however, is considerably more difficult. Any evaluation of the audio summary quality would inherently include the reconstruction quality of the FTO-to-wav model. Furthermore, applying the qualitative evaluation method described in Section 4.5 would require reconstructing the full-length recordings to compare against the skim. For 24h recordings, this implies 120h of computation per recording, amounting to 960h for the entire dataset. While technically feasible, this is computationally expensive and was therefore left for future work.

5.2.3 Robustness of FTO speech privacy enforcement against spectrogram super-resolution algorithms

The fact that FTOs are by design unintelligible was claimed in (Gontier, Lagrange, et al. 2017). They considered an at the time state-of-the-art reconstruction approach combining a) PINV for frequency retrieval and b) the Griffin-Lim algorithm (Griffin et al. 1984) for phase retrieval, and demonstrated that the recovered speech was unintelligible. Based purely on signal processing techniques, this reconstruction method does not consider any *a priori* knowledge on spectro-temporal properties of spoken English. With the reconstruction results provided by FTO-to-wav, we believe that there's a need to re-evaluate the aforementioned claim that is: FTO encoding is *by-design* preserving content privacy.

In order to evaluate potential privacy threats induced by training a Diffspec model within the same pipeline as FTO-to-wav, we categorize out attack models into three

distinct Attack Levels (AL) on FTO data, based on training set selection:

— **AL0** denotes an unintentional attack, occurring when a model is trained on general urban data or without any specialized training. The aim of the attacker in this case is not to recover specifically speech information but rather to reconstruct general audio for analysis purposes. This corresponds to the use of FTO-to-wav.

— **AL1** denotes an attack resulting from training a model on general speech data. In this scenario, the attacker aims to recover speech information from an unknown speaker.

— **AL2** denotes an attack executed by training a model specifically on the voice of a target individual. In this scenario, it is assumed that the attacker has access to a sufficient amounts of clear speech data from his target.

To evaluate the impact of large-scale attacks on sensor data, we anticipate that attackers will utilize Automatic Speech Recognition (ASR) systems on FTO data. Given the uncertainty surrounding the effectiveness of ASRs in processing severely downsampled acoustic signals such as FTOs, we perform a subjective assessment through Human Speech Recognition (HSR) evaluation. This subjective evaluation provides a reference for interpreting ASRs performance under those kind of adverse conditions. To assess the privacy threat potentially induced by the generated audio samples, we measure the Word Error Rate (WER). The WER is a measure of the discrepancy between the reference transcriptions and those produced by HSR or ASR systems on the reconstructed speech. It is calculated as the sum of the number of substitutions, deletions, and insertions required to convert the inferred text into the reference text, divided by the total number of words in the reference text.

For each Attack Level (AL), we select a specific audio dataset to compute FTO data to train a Diffspec model. The chosen datasets are:

— **AL0** dataset: TAU Urban Acoustic Scenes 2020 Mobile dataset (Mesaros, Heittola, and Virtanen 2018), used for training the Diffspec of FTO-to-wav in Section 5.2.1. While it includes some distant voice samples, it primarily focuses on ambient urban sounds, and totals 64 hours of audio.

— **AL1** dataset: Librispeech (Panayotov et al. 2015). It consist of read audiobooks with more than 2,000 speakers. Specifically we use the "train-clean-100" subset, which includes 100 hours of audio data.

— **AL2** dataset: LJSpeech [2]. It comprises 13,000 audio clips from a single speaker reading seven non-fiction books, totalling 24h of audio. Like Librispeech, readings are available through the LibriVox project. The training dataset we considered comprises 12,900 audio clips, as 100 are kept for evaluation.

The pipeline presented in Figure 5.2, applied across different Attack Levels (AL) for training Diffspec, yields in a minimum Word Error Rate (WER) of 93% for AL0, 89% for AL1, and 46% for AL2 for several ASR systems on our LJSpeech test set. More details on those results are available at (Tailleur, Benaatia, et al. 2025) [3].

The Attack Level AL0 performs only slightly better than White Noise, which suggests that using FTO-to-wav is generally safe for urban sound monitoring from a privacy perspective. Results on AL1 and AL2 also indicate a very low risk of extracting intelligible speech information when Diffspec is trained on speech data. However, even with the seemingly high WERs for AL1 and AL2, the risk associated with these attacks is highly context-dependent.

Future work could therefore consider perceptual assessment of acceptable risk thresholds for WER in specific application settings from public spaces, to offices or bedrooms. In the meantime, we consider that recording FTO in urban sensor networks remains privacy-preserving, as these environments are typically noisy and recordings are made from elevated positions, making it very unlikely to capture isolated voice samples. However, we recommend not making FTO recordings from bedrooms or offices publicly available, in order to minimize any potential risk in the future. Note that this evaluation focuses solely on speech content intelligibility. A comprehensive assessment of privacy risks should also include speaker identification and emotion recognition.

5.3 Towards standardization of sound source presence prediction from acoustic measurements

One current limitation to the broader use of PANN-FTO in supporting soundscape-centered urban planning is the absence of any formal legal or normative framework for estimating sound source presence. ISO/DIS 12913-2 (International Standard 2018) suggests that part of soundscape evaluation can rely on the presence of sound sources, based

2. LJSpeech dataset available at: https://keithito.com/LJ-Speech-Dataset/
3. Code operating solely on public data and audio examples are available at: https://modantailleur.github.io/paperThirdOctavePrivacy/

on an established taxonomy (Brown et al. 2011). The standard mentions that this presence may be assessed through questionnaires. Nonetheless, there is no standardized method for assessing the presence of sound sources through automatic classifiers such as PANN-FTO.

The lack of standardization poses a significant barrier to wider deployment, as standardization often plays a key role in establishing the reliability and credibility of measurement tools. We anticipate that future research and standardization efforts will help define how sound source presence should be measured and interpreted in automated systems from a soundscape perspective.

5.4 Rethinking soundscape assessment: the role of non-acoustic factors

Multiple studies over the past decades have shown that soundscape perception is strongly influenced by non-acoustic factors such as visual cues, smells, and contextual information. For instance, (Yong Jeon et al. 2011) demonstrated that both lighting and visual scenery significantly shape how soundscapes are perceived. Similarly, (D'Alessandro et al. 2018) highlighted the role of visual and climatic conditions in shaping listeners' responses to acoustic environments. (X. Xu et al. 2023) and (Zhong et al. 2022) have found that the visible presence of roads, buildings, crowds, sky, or water significantly change the perception of environmental sound. These studies reinforce the idea that relying solely on acoustic measurements may not be sufficient to assess the quality of a soundscape. Multisensory and contextual factors play a central role in how sound is perceived and evaluated, and should therefore be considered in future efforts to visualize and communicate soundscapes.

Sound source prediction could also benefit from incorporating contextual information such as topography or urban infrastructure data that describe where a sensor is deployed. For instance, if a sensor detects church bells in an area where no church is present, this likely indicates a misclassification. By integrating location-aware metadata, such as land use, known landmarks, or building functions into the prediction pipeline, models could filter out ambiguous detections and improve overall accuracy.

The experience of sound is also strongly influenced by psychological phenomena. As shown in (Steffens et al. 2015), the average assessment of a soundscape experience does not necessarily match the *a posteriori* evaluation of it. Listeners are influenced by the peak–end effect, which suggests that people's overall judgments are disproportionately

shaped by the most intense moments and the final part of the experience. Such effects could be incorporated into the prediction pipeline by accounting for both the saliency of sounds and their temporal structure.

5.5 Public communication on soundscape

In Chapter 3, we proposed a visualization framework specifically for urban planning. Extending this approach to the general public involves rethinking how soundscapes are represented, with the goal of making them both accessible and engaging to non-experts.

Traditional psychoacoustic parameters—such as loudness, fluctuation strength, roughness, sharpness, and pitch strength—have shown limited applicability in environmental acoustics. This is largely due to the polyphonic nature of environmental soundscapes, which are often too complex to be meaningfully characterized using such isolated descriptors (Zwicker et al. 2013). As a result, alternative frameworks for developing dedicated soundscape indices have been proposed and explored, notably in (Aletta et al. 2016; J. Kang et al. 2019).

Still, the development of these indices typically involves measuring how individuals perceive a soundscape, which is not the same as helping them understand or grasp what the soundscape actually contains. As a result, most soundscape indices and descriptors are primarily designed for analysis by researchers and experts in the field, and are not necessarily well-suited for communicating soundscape characteristics to the general public. For example, high-level perceptual attributes such as the time of presence of sound sources, which are well-established in soundscape research, were found to be somewhat still unintuitive in interviews with non-experts (Tailleur, Aumond, Tourre, et al. 2024). While providing audio skims, as proposed in Chapter 4, could offer a more concrete representation, this approach remains limited in practice for in-depth analysis, as listening to audio is cognitively demanding and time-consuming compared to visual stimuli.

Future work could explore making sound source presence more accessible through dual representations that combine audio playback with explicit sound source information. For example, based on our work on audio skim generation, we propose audio-visual representations [4] that not only allow users to listen to the soundscape but also display a word map of the detected sound sources. We hypothesize that by first exposing users to the audio

4. Audio-visual representations of audio skims available at:https://modantailleur.github.io/paperAudioSummary/

stimuli, they become familiar with the soundscape and can later rely solely on the word map or on simple source presence scores to recall its acoustic characteristics.

Chapter 6

Conclusion

This thesis explored new approaches to soundscape analysis using long-term acoustic measurements, both in the form of audio waveforms and spectro-temporal representations known as fast third-octave (FTO).

We first developed a novel method to predict the time of presence of traffic, voices, and birds from FTOs, using pre-trained classification models (PANNs) combined with a transcoder that converts FTO data into fine-grained Mel spectrograms. We demonstrated that the predictions made by our model were correlated with perceptual assessments of source presence. On FTO recordings from an urban sensor network, the model successfully retrieved coherent information on sound sources, confirming its practical applicability. We showed that this approach could be extended to detect additional sound sources, such as music, sirens, and church bells, indicating potential for broader source presence prediction tasks.

Building on these predictions, we applied them to the CENSE Lorient dataset to construct an urban analytics tool. This visual analytics framework is designed to support soundscape-oriented urban planning by enabling the exploration of acoustic sensor network data. It guides users through a three-step analytical workflow via interactive multi-source visualizations: (A) polar area diagrams for spatial comparison, (B) geo-referenced 3D data vases linking spatial and temporal patterns, and (C) 2D data vases for fine-grained temporal analysis. The framework enables urban planners to investigate spatio-temporal variations in sound source presence and to assess the effects of soundscape design (SSD) interventions. Feedback from domain experts confirmed the tool's value in supporting data-informed decisions in urban sound planning.

To further enhance soundscape understanding, we introduced an audio summarization method that allows to condense a 24h recording into a 1min audio skim. We evaluated the quality of these skims using two complementary metrics: faithfulness, which captures how well the summary captures the average background sound, and highlightness, which reflects how well the skim represents the most salient sound sources. Our method first

segments the recording into fixed time windows and selects representative excerpts using clustering. In a second step, these excerpts are ordered into a final skim, balancing faithfulness and highlightness via a user controllable parameter z. We showed that the resulting skims provide a strong compromise between the two metrics, and that the z parameter enables users to control well the trade-off based on analysis goals.

This thesis presented new methodologies for soundscape analysis using long-term acoustic data. By proposing tools for source presence estimation, spatio-temporal exploration, and summarization, we contribute to a more actionable and human-centered understanding of sound environments. We hope this work will support future efforts to better characterize urban soundscapes and guide interventions aimed at enhancing their quality for the benefit of all citizens.

Chapter 7

Contributions

7.1 Scientific publications

7.1.1 Publications and submissions in international journals

Reference: Modan Tailleur, Pierre Aumond, Mathieu Lagrange, and Vincent Tourre (2024), « Sound source classification for soundscape analysis using fast third-octave bands data from an urban acoustic sensor network », *in*: *The Journal of the Acoustical Society of America (JASA)*

Source code:
https://github.com/modantailleur/paperSoundscapeSourceClassification

---

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

Reference: Modan Tailleur, Vincent Tourre, Pierre Amond, Mathieu Lagrange, and Ophélie Moreau (2025), « SoundSourceViz: A Visual Analytics Tool for Soundscape-Oriented Urban Planning with Acoustic Sensor Networks », *in*: *(submitted to Computer Graphics Forum)*

Companion page:
https://github.com/modantailleur/paperSoundSourceViz

We present SoundSourceViz, a visual analytics framework designed to support soundscape-oriented urban planning through the exploration of data collected by acoustic sensor networks. Rather than displaying raw audio data, we estimate the perceptual presence of key sound sources—traffic, voices, and birds—using audio classification models, providing interpretable proxies for mechanical, social, and natural components of the urban soundscape. SoundSourceViz uses a three-step analytical workflow with interactive multi-source visualizations: (A) polar area diagrams for spatial comparison, (B) geo-referenced 3D data vases for linking spatial location and temporal patterns, and (C) 2D data vases for precise temporal comparisons. The framework allows urban planners to examine spatial and temporal variations in the presence of multiple sound sources, and to compare conditions before and after soundscape design (SSD) interventions. Feedback from domain experts highlights the value of our approach for supporting data-informed decisions in urban sound planning.

Reference: Modan Tailleur, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2025a), « Audio Skim Generation for Environmental Audio Recordings », *in*: *(to be submitted)*

Companion page:
https://modantailleur.github.io/paperAudioSummary/

Environmental audio recordings offer rich information for soundscape analysis but are often too lengthy for practical listening. To address this, we introduce an automatic method for generating audio summaries, which we call audio skims, designed to highlight key characteristics of a full-length audio. Our approach balances two

objectives: faithfulness, which reflects how closely the summary preserves the background sound of the full-length audio and its evolution throughout time, and highlightness, which reflects how well the summary represents the most salient events of the full-length audio. We define a controllable parameter z to modulate this trade-off and propose a two-stage audio skim generation pipeline: Main Events Identification (MEI) via K-means clustering of audio embeddings, and Sequencing of Main Events (SME) using a greedy algorithm optimized for the z-weighted balance between highlightness and faithfulness. We validate our method on a dataset of 24-hour urban recordings by generating 1-minute summaries. Through both objective and subjective evaluations, we show that our method produces high-quality audio skims and enables controllable trade-offs between faithfulness and highlightness. This approach supports practical and customizable access to environmental recordings, with applications in soundscape research, urban planning, and public communication.

7.1.2 Publications and submissions in international conferences

First author

Reference: Modan Tailleur, Chaymae Benaatia, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2025), « Diffusion-based spectral super-resolution of third octave acoustic sensor data: is privacy at risk? », *in*: *European Signal Processing Conference (EUSIPCO)*

Companion page:
https://modantailleur.github.io/paperThirdOctavePrivacy/

Third octave spectral recording of acoustic sensor data is an effective way of measuring the environment. While there is strong evidence that slow (1s frame, 1 Hz rate) and fast (125ms frame, 8Hz rate) versions lead by-design to unintelligible speech if reconstructed, the advent of high quality reconstruction methods based on diffusion may pose a threat, as those approaches can embed a significant amount of *a priori* knowledge when learned over extensive speech datasets.
This paper aims to assess this risk at three levels of attacks with a growing level of *a priori* knowledge considered at the learning of the diffusion model, a) none, b)

multi-speaker data excluding the target speaker and c) target speaker. Without any prior regarding the speech profile of the speaker (levels a and b), the word-error-rate both for humans and automatic recognition remains higher than 89%.

Reference: Modan Tailleur, Pierre Aumond, Vincent Tourre, and Mathieu Lagrange (2024), « Towards better visualizations of urban sound environments: insights from interviews », *in*: *Internoise*

Urban noise maps and noise visualizations traditionally provide macroscopic representations of noise levels across cities. However, those representations fail at accurately gauging the sound perception associated with these sound environments, as perception highly depends on the sound sources involved. This paper aims at analyzing the need for the representations of sound sources, by identifying the urban stakeholders for whom such representations are assumed to be of importance. Through spoken interviews with various urban stakeholders, we have gained insight into current practices, the strengths and weaknesses of existing tools and the relevance of incorporating sound sources into existing urban sound environment representations. Three distinct use of sound source representations emerged in this study: 1) noise-related complaints for industrials and specialized citizens, 2) soundscape quality assessment for citizens, and 3) guidance for urban planners. Findings also reveal diverse perspectives for the use of visualizations, which should use indicators adapted to the target audience, and enable data accessibility.

Reference: Modan Tailleur, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2025b), *Enforcing Speech Content Privacy in Environmental Sound Recordings using Segment-wise Waveform Reversal*, DOI: 10.48550/arXiv.2507.08412

Companion page:
https://modantailleur.github.io/paperSpeechContentPrivacyEnforcement/

Environmental sound recordings often contain intelligible speech, raising privacy concerns that limit analysis, sharing and reuse of data. In this paper, we introduce

a method that renders speech unintelligible while preserving both the integrity of the acoustic scene, and the overall audio quality. Our approach involves reversing waveform segments to distort speech content. This process is enhanced through a voice activity detection and speech separation pipeline, which allows for more precise targeting of speech.
In order to demonstrate the effectivness of the proposed approach, we consider a three-part evaluation protocol that assesses: 1) speech intelligibility using Word Error Rate (WER), 2) sound sources detectability using Sound source Classification Accuracy-Drop (SCAD) from a widely used pre-trained model, and 3) audio quality using the Fréchet Audio Distance (FAD), computed with our reference dataset that contains unaltered speech. Experiments on this simulated evaluation dataset, which consists of linear mixtures of speech and environmental sound scenes, show that our method achieves satisfactory speech intelligibility reduction (97.9% WER), minimal degradation of the sound sources detectability (2.7% SCAD), and high perceptual quality (FAD of 1.40). An ablation study further highlights the contribution of each component of the pipeline. We also show that incorporating random splicing to our speech content privacy enforcement method can enhance the algorithm's robustness to attempt to recover the clean speech, at a slight cost of audio quality.

Co-author

Reference: Dick Botteldooren, Nele De Poortere, Modan Tailleur, and Arnaud Can (2024), « Characterizing urban acoustic environments by clustering sensor node measurements », *in*: *INTERNOISE*

Equivalent sound pressure levels have been widely used to assess the impact of the sonic environment on persons. At the same time, it has been demonstrated that sound sensor networks allow to extract a wealth of information about the living environment. Nevertheless it has not been clearly demonstrated how to summarize this information. To this end, we used advanced non-linear clustering methods to group 15-minute sound environments based on a wide range of classical acoustic indicators such as peak levels, statistical levels, spectral center of gravity. Machine learning methods identifying the sound sources that are present based on 1/3 octave

band levels measured at 125 msec interval, are then used to better understand why these clusters emerge. The methodology is applied to 150 bedroom window measurements gathered in the Equal-Life project. Preliminary results show that groups of pre-school children with specific cognitive development and mental health are related to different diurnal patterns of sound environment clusters. These results can be interpreted in the larger context of living environments extracted from environmental sound sensing networks.

7.1.3 Publications in international workshops

Reference: Modan Tailleur, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2023), « Spectral trancoder: using pretrained urban sound classifiers on undersampled spectral representations », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*

Source code:
https://github.com/modantailleur/paperSpectralTranscoder

Slow or fast third-octave bands representations (with a frame resp. every 1-s and 125-ms) have been a de facto standard for urban acoustics, used for example in long-term monitoring applications. It has the advantages of requiring few storage capabilities and of preserving privacy. As most audio classification algorithms take Mel spectral representations with very fast time weighting (ex. 10-ms) as input, very few studies have tackled classification tasks using other kinds of spectral representations of audio such as slow or fast third-octave spectra.
In this paper, we present a convolutional neural network architecture for transcoding fast third-octave spectrograms into Mel spectrograms, so that it could be used as input for robust pre-trained models such as YAMNet or PANN models. Compared to training a model that would take fast third-octave spectrograms as input, this approach is more effective and requires less training effort. Even if a fast third-octave spectrogram is less precise both on time and frequency dimensions, experiments show that the proposed method still allows for classification accuracy of 62.4% on UrbanSound8k and 0.44 macro AUPRC on SONYC-UST.

Reference: Modan Tailleur, Vincent Lostanlen, Jean-Philippe Rivière, and Pierre Aumond (2024), « Machine listening in a neonatal intensive care unit », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*

Oxygenators, alarm devices, and footsteps are some of the most common sound sources in a hospital. Detecting them has scientific value for environmental psychology but comes with challenges of its own: namely, privacy preservation and limited labeled data. In this paper, we address these two challenges via a combination of edge computing and cloud computing. For privacy preservation, we have designed an acoustic sensor which computes third-octave spectrograms on the fly instead of recording audio waveforms. For sample-efficient machine learning, we have repurposed a pretrained audio neural network (PANN) via spectral transcoding and label space adaptation. A small-scale study in a neonatological intensive care unit (NICU) confirms that the time series of detected events align with another modality of measurement: i.e., electronic badges for parents and healthcare professionals. Hence, this paper demonstrates the feasibility of polyphonic machine listening in a hospital ward while guaranteeing privacy by design.

7.2 Open-source libraries

Name: FAST-TO-EARS

Link:
https://github.com/modantailleur/fast-to-ears

FAST-TO-EARS enables sound source predictions using fast third-octave recorded datasets. Utilizing the transcoding method introduced in (Tailleur, Lagrange, et al. 2023), it transcodes the third-octave spectro-temporal representations into fine-grained Mel spectrograms. These spectrograms then serve as input for PANNs (Kong et al. 2020), which are large-scale pre-trained classifiers capable of predicting the presence of over 527 different sound sources.

Name: FAST-TO-WAV

Link:
https://github.com/modantailleur/fast-to-wav

FAST-TO-WAV provides means to transform fast third-octave data into high-quality audio waveforms. The model is based on the diffusion architecture described in (Tailleur, Benaatia, et al. 2025). It is trained on the TAU Urban Acoustic Scenes 2020 Mobile, Development dataset (Mesaros, Heittola, and Virtanen 2018).

7.3 Dataset

Name: CitySpeechMix: A Simulated Dataset of Speech and Urban Sound Mixtures from LibriSpeech and SONYC-US

Link:
https://zenodo.org/records/15405950

CitySpeechMix is a simulated audio dataset that mixes speech excerpts from LibriSpeech (Panayotov et al. 2015) with environmental recordings from SONYC-UST (Bello et al. 2019) to create controlled mixtures of voice and background noise. Each audio file is accompanied by the corresponding LibriSpeech transcription and the SONYC-UST sound class labels. A mapping is also provided between the selected SONYC-UST sound classes and their corresponding AudioSet (Gemmeke et al. 2017) categories.

Appendix A

Aggregation algorithm for the display of the polar area diagrams network

To mitigate visual clutter in dense glyph-based visualizations, McNabb et al. proposed a hierarchical glyph aggregation method (McNabb and Laramee 2019; McNabb, Laramee, and M. Wilson 2018). While effective in contexts where data can be grouped into meaningful spatial regions (e.g., U.S. states or administrative areas), this approach is not directly applicable to the use case of Chapter 3. In our scenario, sensors are positioned at discrete point locations without associated spatial regions, making region-based aggregation unsuitable.

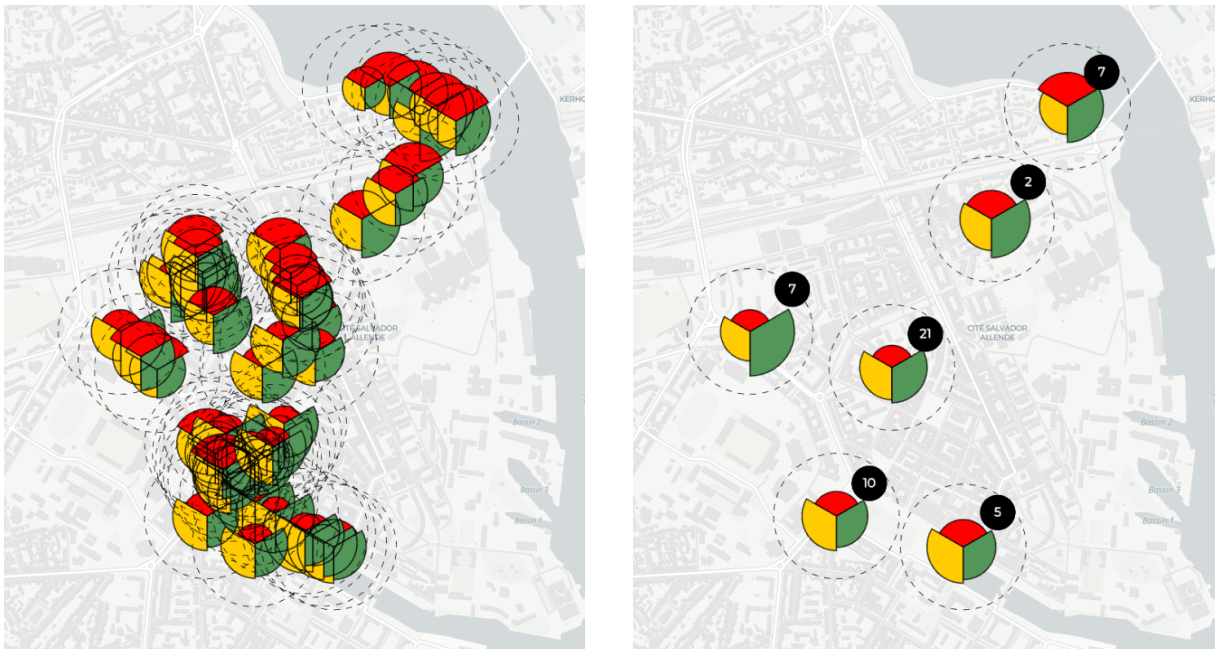
Instead, we implement a greedy aggregation algorithm that removes overlapping polar area diagrams at each zoom level. Two diagrams are considered to overlap when their dashed boundary circles intersect. Let r denote the radius of such a circle. For each overlapping pair, we compute the intersection area of the corresponding circles. When the distance between their centers is $d < 2r$, the intersection area is given by:

$$A = 2r^2 \cos^{-1}\left(\frac{d}{2r}\right) - \frac{d}{2}\sqrt{4r^2 - d^2} \tag{A.1}$$

The aggregation algorithm proceeds greedily: at each step, it selects the glyph with the largest sum of intersection area with other glyphs. This glyph is retained, and all overlapping glyphs that intersect with it are removed. The process repeats until either only one glyph remains or no overlaps are left.

The algorithm is computed across 100 discrete zoom levels, from the lowest zoom level (where only one glyph is shown) to the highest (where all glyphs are displayed). Since each glyph represents the position of a sensor, we prioritize continuity during zooming. Specifically, we ensure that any glyph displayed at a given zoom level remains visible at

higher zoom levels. At each new zoom step, we thus retain all previously preserved glyphs and discard any new glyphs that would overlap with them. The greedy algorithm is then applied to the remaining candidates. The result of this aggregation method is shown in Figure A.1.



(a) Without aggregation

(b) With aggregation

Figure A.1 – Polar area diagram network without and with glyph aggregation. In (b), the number of sensors hidden beneath a glyph is specified within a black circle.

BIBLIOGRAPHY

Abeßer, Jakob, Marco Götze, Tobias Clauß, Dominik Zapf, Christian Kühn, Hanna Lukashevich, Stephanie Kühnlenz, and Stylianos I. Mimilakis (2019), « Urban noise monitoring in the stadtlärm project-a field report », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Aiello, Luca Maria, Rossano Schifanella, Daniele Quercia, and Francesco Aletta (2016), « Chatty maps: constructing sound maps of urban areas from social media data », *in*: *Royal Society Open Science*.

Alaybek, Balca, Reeshad S. Dalal, Shea Fyffe, John A. Aitken, You Zhou, Xiao Qu, Alexis Roman, and Julia I. Baines (2022), « All's well that ends (and peaks) well? A meta-analysis of the peak-end rule and duration neglect », *in*: *Organizational Behavior and Human Decision Processes*.

Aletta, Francesco, Jian Kang, and Osten Axelsson (2016), « Soundscape descriptors and a conceptual framework for developing predictive soundscape models », *in*: *Landscape and Urban Planning*.

Apostolidis, Evlampios, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras (2020), « AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization », *in*: *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*.

— (2021), « Video summarization using deep neural networks: A survey », *in*: *Proceedings of the IEEE*.

Apostolidis, Evlampios, Alexandros I. Metsai, Eleni Adamantidou, Vasileios Mezaris, and Ioannis Patras (2019), « A Stepwise, Label-based Approach for Improving the Adversarial Training in Unsupervised Video Summarization », *in*: *International Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV)*.

Argaw, Dawit Mureja, Mattia Soldan, Alejandro Pardo, Chen Zhao, Fabian Caba Heilbron, Joon Son Chung, and Bernard Ghanem (2024), « Towards automated movie trailer generation », *in*: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Arthur, David and Sergei Vassilvitskii (2006), *k-means++: The advantages of careful seeding*, tech. rep., Stanford.

Asdrubali, Francesco, Francesco D'Alessandro, Giorgio Baldinelli, and Brigitte Schulte-Fortkamp (2014), « From the soundscape to the architectural redevelopment of an outdoor public space », *in*: *Forum Acusticum*.

Aspuru, Itziar, Igone García, Laura Gutierrez, Karmele Herranz, and Juan Angel Acero (2014), « Comfort Urban Places: A Proposal for Designing Unique Places that Provide Welfare. Practical Case: Municipality of Sestao », *in*: *Smart City Expo World Congress (SCEWC)*.

Aumond, Pierre, Arnaud Can, Bert De Coensel, Dick Botteldooren, Carlos Ribeiro, and Catherine Lavandier (2017), « Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context », *in*: *Acta Acustica*.

Aumond, Pierre, Arnaud Can, Vivien Mallet, Bert De Coensel, Carlos Ribeiro, Dick Botteldooren, and Catherine Lavandier (2018), « Kriging-based spatial interpolation from measurements for sound level mapping in urban areas », *in*: *The Journal of the Acoustical Society of America (JASA)*.

Axelsson, Osten, Mats E. Nilsson, and Birgitta Berglund (2010), « A principal components model of soundscape perception », *in*: *The Journal of the Acoustical Society of America (JASA)*.

Babisch, W., H. Fromme, A. Beyer, and H. Ising (2001), « Increased catecholamine levels in urine in subjects exposed to road traffic noise: the role of stress hormones in noise research », *in*: *Environment international*.

Bach, B., P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale (2017), « A Descriptive Framework for Temporal Data Visualizations Based on Generalized Space-Time Cubes », *in*: *Computer Graphics Forum (CGF)*.

Bansal, Anam and Naresh Kumar Garg (2022), « Environmental Sound Classification: A descriptive review of the literature », *in*: *Intelligent Systems with Applications (ISWA)*.

Bello, Juan P., Claudio Silva, Oded Nov, R. Luke Dubois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy (2019), « SONYC: a system for monitoring, analyzing, and mitigating urban noise pollution », *in*: *Communications of the ACM*.

Berman, Marc G., John Jonides, and Stephen Kaplan (2008), « The Cognitive Benefits of Interacting With Nature », *in*: *Psychological Science*.

Bild, Edda, Matt Coler, Karin Pfeffer, and Luca Bertolini (2016), « Considering Sound in Planning and Designing Public Spaces: A Review of Theory and Applications and a Proposed Framework for Integrating Research and Practice », *in*: *Journal of Planning Literature (JPL)*.

Botteldooren, Dick, Nele De Poortere, Modan Tailleur, and Arnaud Can (2024), « Characterizing urban acoustic environments by clustering sensor node measurements », *in*: *INTERNOISE*.

Botteldooren, Dick, Catherine Lavandier, Anna Preis, Daniele Dubois, Itziar Aspuru, Catherine Guastavino, Lex Brown, Mats Nilsson, and Tjeerd C. Andringa (2011), « Understanding urban and natural soundscapes », *in*: *Forum Acusticum*.

Brewer, Cynthia A. (1999), « Color use guidelines for data representation », *in*: *Section on statistical graphics*.

Brown, Alan Lex, Jian Kang, and Truls Gjestland (2011), « Towards standardization in soundscape preference assessment », *in*: *Applied acoustics*.

Burges, Christopher JC, John C. Platt, Daniel Plastina, Erin L. Renshaw, and Henrique S. Malvar (2008), « Systems and methods for generating audio thumbnails ».

Buxton, Rachel T., Amber L. Pearson, Claudia Allou, Kurt Fristrup, and George Wittemyer (2021), « A synthesis of health benefits of natural sounds and their distribution in national parks », *in*: *The Proceedings of the National Academy of Sciences (PNAS)*.

Can, Arnaud, Pierre Aumond, Bert De Coensel, Carlos Ribeiro, Dick Botteldooren, and Catherine Lavandier (2018), « Probabilistic modelling of the temporal variability of urban sound levels », *in*: *Acta Acustica*.

Can, Arnaud, Judicaël Picaut, Jeremy Ardouin, Pierre Crepeaux, Erwan Bocher, David Ecotiere, and Mathieu Lagrange (2021), « CENSE Project: general overview », *in*: *European Congress on Noise Control Engineering (EuroNoise)*.

Cartwright, Mark, Ana Elisa Mendez Mendez, Jason Cramer, Vincent Lostanlen, Graham Dove, Ho-Hsiang Wu, Justin Salamon, Oded Nov, and Juan Bello (2019), « SONYC urban sound tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Chen, Ke, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov (2022), « HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection », *in*: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chen, Sanyuan, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei (2022), *BEATs: Audio Pre-Training with Acoustic Tokenizers*, DOI: 10.48550/arXiv.2212.09058.

Chen, Yi-Syuan, Yun-Zhu Song, and Hong-Han Shuai (2022), « Spec: Summary preference decomposition for low-resource abstractive summarization », *in*: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.

Chen, Xiuying, Shen Gao, Mingzhe Li, Qingqing Zhu, Xin Gao, and Xiangliang Zhang (2024), « Write Summary Step-by-Step: A Pilot Study of Stepwise Summarization », *in*: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.

Chirigati, Fernando, Harish Doraiswamy, Theodoros Damoulas, and Juliana Freire (2016), « Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets », *in*: *ACM International Conference on Management of Data (MOD)*.

Cooper, Matthew and Jonathan Foote (2002), « Automatic Music Summarization via Similarity Analysis. », *in*: *The International Society for Music Information Retrieval (ISMIR)*.

— (2003), « Summarizing popular music via structural similarity analysis », *in*: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

D'Alessandro, Francesco, Luca Evangelisti, Claudia Guattari, Gianluca Grazieschi, and Federico Orsini (2018), « Influence of visual aspects and other features on the soundscape assessment of a university external area », *in*: *Building Acoustics*.

De Coensel, Bert, Annelies Bockstael, Luc Dekoninck, Dick Botteldooren, Brigitte Schulte-Fortkamp, Jian Kang, and Mats E. Nilsson (2010), « The soundscape approach for early stage urban planning: a case study », *in*: *Internoise*.

Deng, Zikun, Di Weng, Jiahui Chen, Ren Liu, Zhibin Wang, Jie Bao, Yu Zheng, and Yingcai Wu (2019), « AirVis: Visual analytics of air pollution propagation », *in*: *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Dove, Graham, Charlie Mydlarz, Juan Pablo Bello, and Oded Nov (2022), « Sounds of New York city », *in*: *Interactions*.

Ellison, William D., Alec C. Trahan, Joseph C. Pinzon, Megan E. Gillespie, Logan M. Simmons, and Kendel Y. King (2020), « For whom, and for what, is experience sampling more accurate than retrospective report? », *in*: *Personality and Individual Differences*.

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

European Environment Agency (2025), *Environmental noise in Europe - 2025*, tech. rep., European Environment Agency.

Farrés, Júlia Camps (2015), « Barcelona noise monitoring network », *in*: *European Congress on Noise Control Engineering (EuroNoise)*.

Ferreira, Nivan, Marcos Lage, Harish Doraiswamy, Huy Vo, Luc Wilson, Heidi Werner, Muchan Park, and Cláudio Silva (2015), « Urbane: A 3D framework to support data driven decision making in urban development », *in*: *IEEE conference on Visual Analytics Science and Technology (VAST)*.

Ferreira, Nivan, Jorge Poco, Huy T. Vo, Juliana Freire, and Cláudio T. Silva (2013), « Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips », *in*: *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Furui, Sadaoki, Tomonori Kikuchi, Yosuke Shinnaka, and Chiori Hori (2004), « Speech-to-text and speech-to-speech summarization of spontaneous speech », *in*: *IEEE Transactions on Speech and Audio Processing (TASP)*.

Gautier, Jacques, Maria-Jesus Lobo, Benjamin Fau, Armand Drugeon, Sidonie Christophe, and Guillaume Touya (2021), « COVID-19 geoviz for spatio-temporal structures detection », *in*: *Proceedings of the International Cartographic Association (ICA)*.

Gemmeke, Jort F., Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter (2017), « Audio set: An ontology and human-labeled dataset for audio events », *in*: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Gomez, Saúl, Catherine Domínguès, Pierre Aumond, Catherine Lavandier, Gaëtan Palka, and Kamal Serrhini (2017), « Cartographic Representation of Soundscape: Proposals and Assessment », *in*: *Citizen Empowered Mapping*, ed. by Michael Leitner and Jamal Jokar Arsanjani.

Gontier, Félix (2020), « Analysis and synthesis of urban sound scenes using deep learning techniques », PhD Thesis, École centrale de Nantes.

Gontier, Félix, Mathieu Lagrange, Pierre Aumond, Arnaud Can, and Catherine Lavandier (2017), « An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach », *in*: *Sensors*.

Gontier, Félix, Catherine Lavandier, Pierre Aumond, Mathieu Lagrange, and Jean-François Petiot (2019), « Estimation of the perceived time of presence of sources in urban acoustic environments using deep learning techniques », *in*: *Acta Acustica*.

Gontier, Félix, Vincent Lostanlen, Mathieu Lagrange, Nicolas Fortin, Catherine Lavandier, and Jean-François Petiot (2021), « Polyphonic training set synthesis improves self-supervised urban sound classification », *in*: *The Journal of the Acoustical Society of America (JASA)*.

Griffin, Daniel and Jae Lim (1984), « Signal estimation from modified short-time Fourier transform », *in*: *IEEE Transactions on acoustics, Speech, and Signal Processing (TASSP)*.

Guo, Lian-Huan, Shi Cheng, Jiang Liu, Yiyang Wang, Yangsheng Cai, and Xin-Chen Hong (2022), « Does social perception data express the spatio-temporal pattern of perceived urban noise? A case study based on 3,137 noise complaints in Fuzhou, China », *in*: *Applied Acoustics*.

Han, Seungu and Junhyeok Lee (2022), « NU-Wave 2: A General Neural Audio Upsampling Model for Various Sampling Rates », *in*: *Interspeech*.

He, Xufeng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan (2019), « Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks », *in*: *ACM International Conference on Multimedia (ICM)*.

Hellström, Björn, Mats E. Nilsson, Östen Axelsson, and Peter Lundén (2014), « Acoustic design artifacts and methods for urban soundscapes: a case study on the qualitative dimensions of sounds », *in*: *Journal of Architectural and Planning Research (JAPR)*.

Hill, Andrew P., Peter Prince, Evelyn Piña Covarrubias, C. Patrick Doncaster, Jake L. Snaddon, and Alex Rogers (2018), « AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment », *in*: *Methods in Ecology and Evolution*, ed. by Nick Isaac.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean (2015), « Distilling the Knowledge in a Neural Network », *in*: *stat*.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020), « Denoising diffusion probabilistic models », *in*: *International Conference on Neural Information Processing Systems (NeurIPS)*.

Hoffmann, Barbara, Susanne Moebus, Andreas Stang, Eva-Maria Beck, Nico Dragano, Stephan Möhlenkamp, Axel Schmermund, Michael Memmesheimer, Klaus Mann, and Raimund Erbel (2006), « Residence close to high traffic and prevalence of coronary heart disease », *in*: *European heart journal*.

Hong, Joo Young and Jin Yong Jeon (2015), « Influence of urban contexts on soundscape perceptions: A structural equation modeling approach », *in*: *Landscape and Urban Planning*.

— (2017), « Relationship between spatiotemporal variability of soundscape and urban morphology in a multifunctional urban area: A case study in Seoul, Korea », *in*: *Building and Environment*.

Hou, Yuanbo, Qiaoqiao Ren, Huizhong Zhang, Andrew Mitchell, Francesco Aletta, Jian Kang, and Dick Botteldooren (2023), « AI-based soundscape analysis: Jointly identifying sound sources and predicting annoyance », *in*: *The Journal of the Acoustical Society of America (JASA)*.

IEC (2013), *61672-1: 2013 Electroacoustics—Sound Level Meters—Part 1: Specifications*.

International Standard (2014), *ISO/DIS 12913-1, 2014: Acoustics – Soundscape – Part 1: Definition and conceptual framework*.

— (2018), *ISO/DIS 12913-2, 2018: Acoustics – Soundscape – Part 2: Data collection and reporting requirements*.

Itoh, Masahiko, Daisaku Yokoyama, Masashi Toyoda, Yoshimitsu Tomita, Satoshi Kawamura, and Masaru Kitsuregawa (2014), « Visual fusion of mega-city big data: An application to traffic and tweets data analysis of Metro passengers », *in*: *IEEE International Conference on Big Data*.

Jeon, Jin Yong and Joo Young Hong (2015), « Classification of urban park soundscapes through perceptions of the acoustical environments », *in*: *Landscape and urban planning*.

Jiang, Nanzhu and Meinard Müller (2014), « Towards efficient audio thumbnailing », *in*: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Johnston, Kevin, Jay M. Ver Hoef, Konstantin Krivoruchko, and Neil Lucas (2001), *Using ArcGIS geostatistical analyst*, vol. 380, Esri Redlands.

Jung, Yunjae, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon (2019), « Discriminative feature learning for unsupervised video summarization », *in*: *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kang, Jian, Francesco Aletta, Tin Oberman, Mercede Erfanian, Magdalena Kachlicka, Matteo Lionello, and Andrew Mitchell (2019), « Towards soundscape indices », *in*: *International Congress on Acoustics (ICA)*.

Kang, Minsung, Sangshin Oh, Hyeongi Moon, Kyungyun Lee, and Ben Sangbae Chon (2023), « FALL-E: A Foley Sound Synthesis Model and Strategies », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Kawar, Bahjat, Michael Elad, Stefano Ermon, and Jiaming Song (2022), « Denoising diffusion restoration models », *in*: *International Conference on Neural Information Processing Systems (NeurIPS)*.

Kim, Sangho, Sungtak Kim, Suk-bong Kwon, and Hoirin Kim (2006), « A music summarization scheme using tempo tracking and two stage clustering », *in*: *IEEE International Workshop on Multimedia Signal Processing (MMSP)*.

Kong, Qiuqiang, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley (2020), « Panns: Large-scale pretrained audio neural networks for audio pattern recognition », *in*: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.

Kraak, Menno-Jan and Alexandra Koussoulakou (2005), « A Visualization Environment for the Space-Time-Cube », *in*: *Developments in Spatial Data Handling*.

Lavandier, Catherine, Pierre Aumond, Arnaud Can, Félix Gontier, Mathieu Lagrange, and Gwendall Petit (2021), « Urban sensor network for characterizing the sound environment in Lorient (France) through an automatic assessment of traffic, voice and bird presence ratios », *in*: *European Congress on Noise Control Engineering (EuroNoise)*.

Lavandier, Catherine, Pierre Aumond, Saul Gomez, and Catherine Dominguès (2016), « Urban soundscape maps modelled with geo-referenced data », *in*: *Noise Mapping*.

Lavandier, Catherine and Boris Defréville (2006), « The contribution of sound source characteristics in the assessment of urban soundscapes », *in*: *Acta acustica*.

Lee, Jaesung, Jihae Yoon, Seongwon Lee, and Dae-Won Kim (2016), « Supervised music summarization for human-intuitive highlight identification », *in*: *International Journal of Machine Learning and Computing (IJMLC)*.

Lenormand, Maxime, Bruno Gonçalves, Antònia Tugores, and José J. Ramasco (2015), « Human diffusion and city influence », *in*: *Journal of The Royal Society Interface (JRSI)*.

Li, Jing, Jean-Bernard Martens, and Jarke J. Van Wijk (2010), « A model of symbol size discrimination in scatterplots », *in*: *ACM Conference on Human Factors in Computing Systems (CHI)*.

Lichtenfeld, Stephanie, Andrew J. Elliot, Markus A. Maier, and Reinhard Pekrun (2012), « Fertile Green: Green Facilitates Creative Performance », *in*: *Personality and Social Psychology Bulletin*.

Licitra, G., L. Brusci, and M. Cobianchi (2010), « Italian Sonic Gardens: An Artificial Soundscape Approach for New Action Plans », *in*: *Designing Soundscape for Sustainable Urban Development*.

Liu, Jiang, Jian Kang, Tao Luo, Holger Behm, and Timothy Coppack (2013), « Spatiotemporal variability of soundscapes in a multiple functional urban area », *in*: *Landscape and urban planning*.

Liu, Qing, Lei Chen, Yuan Yuan, and Huarui Wu (2021), « History reuse and bag-of-words loss for long summary generation », *in*: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019), *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, DOI: 10.48550/arXiv.1907.11692.

Logan, Beth and Stephen Chu (2000), « Music summarization using key phrases », *in*: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

MacQueen, James (1967), « Some methods for classification and analysis of multivariate observations », *in*: *Berkeley Symposium on Mathematical Statistics and Probability (BSMSP)*.

Mahasseni, Behrooz, Michael Lam, and Sinisa Todorovic (2017), « Unsupervised video summarization with adversarial lstm networks », *in*: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Malik, Abish, Ross Maciejewski, Niklas Elmqvist, Yun Jang, David S. Ebert, and Whitney Huang (2012), « A correlative analysis process in a visual analytics environment », *in*: *IEEE Conference on Visual Analytics Science and Technology (VAST)*.

McNabb, Liam and Robert S. Laramee (2019), « Multivariate maps—a glyph-placement algorithm to support multivariate geospatial visualization », *in*: *Information*.

McNabb, Liam, Robert S. Laramee, and Max Wilson (2018), « When Size Matters: Towards Evaluating Perceivability of Choropleths », *in*: *Computer Graphics and Visual Computing (CGVC)*.

Mesaros, Annamaria, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen (2017), « DCASE 2017 chal-

lenge setup: Tasks, datasets and baseline system », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen (2018), « A multi-device dataset for urban acoustic scene classification », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Mietlicki, Fanny, Christophe Mietlicki, and Matthieu Sineau (2015), « An Innovative Approach for long term environmental noise measurement: RUMEUR Network in the Paris Region », *in*: *European Congress on Noise Control Engineering (EuroNoise)*.

Miller, George A. (1956), « The magical number seven, plus or minus two: Some limits on our capacity for processing information. », *in*: *Psychological review*.

Miranda, Fabio, Harish Doraiswamy, Marcos Lage, Luc Wilson, Mondrian Hsieh, and Claudio T. Silva (2018), « Shadow accrual maps: Efficient accumulation of city-scale shadows over time », *in*: *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Miranda, Fabio, Harish Doraiswamy, Marcos Lage, Kai Zhao, Bruno Gonçalves, Luc Wilson, Mondrian Hsieh, and Cláudio T. Silva (2016), « Urban pulse: Capturing the rhythm of cities », *in*: *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Miranda, Fabio, Marcos Lage, Harish Doraiswamy, Charlie Mydlarz, Justin Salamon, Yitzchak Lockerman, Juliana Freire, and Claudio T. Silva (2018), « Time Lattice: A Data Structure for the Interactive Visual Analysis of Large Time Series », *in*: *Computer Graphics Forum (CGF)*.

Mitchell, Andrew, Tin Oberman, Francesco Aletta, Magdalena Kachlicka, Matteo Lionello, Mercede Erfanian, and Jian Kang (2021), « Investigating urban soundscapes of the COVID-19 lockdown: A predictive soundscape modeling approach », *in*: *The Journal of the Acoustical Society of America (JASA)*.

Moreno, Juan José Franco, Rosa Maria Alsina-Pages, and Pere Pau Vázquez Alcocer (2022), « Visual analysis of environmental noise data », *in*: *International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing (CGVCVIP)*.

Mueller, Shane T., Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein (2019), *Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI*, DOI: 10.48550/arXiv.1902.01876.

Muller, Meinard, Nanzhu Jiang, and Peter Grosche (2012), « A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing », *in*: *IEEE Transactions on audio, speech, and language processing*.

Nieuwenhuijsen, Arianne N. van (2019), « Music Thumbnailing by Hooks », MA thesis.

Nightingale, Florence (1857), *Mortality of the British Army*, Harrison and Sons.

Nilsson, M., Dick Botteldooren, and Bert De Coensel (2007), « Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas », *in*: *International Congress on Acoustics (ICA)*.

Noulas, Anastasios, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo (2012), « A tale of many cities: universal patterns in human urban mobility », *in*: *PloS one*.

Oldoni, Damiano, Bert De Coensel, Annelies Bockstael, Michiel Boes, Bernard De Baets, and Dick Botteldooren (2015), « The acoustic summary as a tool for representing urban sound environments », *in*: *Landscape and Urban Planning*.

Ondov, Brian, Nicole Jardine, Niklas Elmqvist, and Steven Franconeri (2018), « Face to face: Evaluating visual comparison », *in*: *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Ooi, Kenneth, Karn N. Watcharasupat, Santi Peksi, Furi Andi Karnapi, Zhen-Ting Ong, Danny Chua, Hui-Wen Leow, Li-Long Kwok, Xin-Lei Ng, and Zhen-Ann Loh (2021), « A Strongly-Labelled Polyphonic Dataset of Urban Sounds with Spatiotemporal Context », *in*: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Ortner, Thomas, Johannes Sorger, Harald Steinlechner, Gerd Hesina, Harald Piringer, and Eduard Gröller (2016), « Vis-a-ware: Integrating spatial and non-spatial visualization for visibility-aware urban planning », *in*: *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur (2015), « Librispeech: an asr corpus based on public domain audio books », *in*: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Penrose, Roger (1955), « A generalized inverse for matrices », *in*: *Mathematical proceedings of the Cambridge philosophical society*.

Piczak, Karol J. (2015), « ESC: Dataset for Environmental Sound Classification », *in*: *ACM International Conference on Multimedia (ICM)*.

Pravossoudovitch, Karyn, Francois Cury, Steve G. Young, and Andrew J. Elliot (2014), « Is red the colour of danger? Testing an implicit red–danger association », *in*: *Ergonomics*.

Puyana-Romero, Virginia, Giuseppe Ciaburro, Giovanni Brambilla, Christiam Garzón, and Luigi Maffei (2019), « Representation of the soundscape quality in urban areas through colours », *in*: *Noise Mapping*.

Quercia, Daniele and Diego Saez (2014), « Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use », *in*: *IEEE Pervasive Computing*.

Ricciardi, Paola, Pauline Delaitre, Catherine Lavandier, Francesca Torchia, and Pierre Aumond (2015), « Sound quality indicators for urban places in Paris cross-validated by Milan data », *in*: *The Journal of the Acoustical Society of America (JASA)*.

Robertson, Alan R. (1977), « The CIE 1976 Color-Difference Formulae », *in*: *Color Research & Application*.

Rulff, Joao, Fabio Miranda, Maryam Hosseini, Marcos Lage, Mark Cartwright, Graham Dove, Juan Bello, and Claudio T. Silva (2022), « Urban Rhapsody: Large-scale exploration of urban soundscapes », *in*: *Computer Graphics Forum (CGF)*.

Saharia, Chitwan, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi (2022), « Image super-resolution via iterative refinement », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello (2014), « A dataset and taxonomy for urban sound research », *in*: *ACM International Conference on Multimedia (ICM)*.

Sawata, Ryosuke, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Takashi Shibuya, Shusuke Takahashi, and Yuki Mitsufuji (2023), « Diffiner: A Versatile Diffusion-based Generative Refiner for Speech Enhancement », *in*: *Interspeech*.

Schaffer, Murray (1979), « Le paysage sonore », *in*: *París: Ed. Jean-Claude Lattés*.

Serrà, Joan, Santiago Pascual, Jordi Pons, R. Oguz Araz, and Davide Scaini (2022), *Universal Speech Enhancement with Score-based Diffusion*, DOI: 10.48550/arXiv.2206.03065.

Sevillano, Xavier, Joan Claudi Socoró, Francesc Alías, Patrizia Bellucci, Laura Peruzzi, Simone Radaelli, Paola Coppi, Luca Nencini, Andrea Cerniglia, Alessandro Bisceglie, Roberto Benocci, and Giovanni Zambon (2016), « DYNAMAP – Development of low cost sensorsnetworks for real time noise mapping », *in*: *Noise Mapping*.

Shao, Xi, Namunu C. Maddage, Changsheng Xu, and Mohan S. Kankanhalli (2005), « Automatic music summarization based on music structure analysis », *in*: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Shneiderman, Ben (2003), « The eyes have it: A task by data type taxonomy for information visualizations », *in*: *The craft of information visualization*.

Siliezar, Jonathan, Arnaud Can, Pierre Aumond, and Paul Chapron (2024), « Mapping human voice in urban spaces », *in*: *Internoise*.

Song, Xinhui, Li Sun, Jie Lei, Dapeng Tao, Guanhong Yuan, and Mingli Song (2016), « Event-based large scale surveillance video summarization », *in*: *Neurocomputing*.

Stansfeld, Stephen A., Birgitta Berglund, Charlotte Clark, Isabel Lopez-Barrio, Peter Fischer, Evy Öhrström, Mary M. Haines, Jenny Head, Staffan Hygge, and Irene Van Kamp (2005), « Aircraft and road traffic noise and children's cognition and health: a cross-national study », *in*: *The Lancet*.

Steffens, Jochen and Catherine Guastavino (2015), « Trend effects in momentary and retrospective soundscape judgments », *in*: *Acta Acustica*.

Sueur, Jerome, Thierry Aubin, and Caroline Simonis (2008), « Seewave, a free modular tool for sound analysis and synthesis », *in*: *Bioacoustics*.

Sundaram, Hari, Lexing Xie, and Shih-Fu Chang (2002), « A utility framework for the automatic generation of audio-visual skims », *in*: *ACM International Conference on Multimedia (ICM)*.

Tailleur, Modan, Pierre Aumond, Mathieu Lagrange, and Vincent Tourre (2024), « Sound source classification for soundscape analysis using fast third-octave bands data from an urban acoustic sensor network », *in*: *The Journal of the Acoustical Society of America (JASA)*.

Tailleur, Modan, Pierre Aumond, Vincent Tourre, and Mathieu Lagrange (2024), « Towards better visualizations of urban sound environments: insights from interviews », *in*: *Internoise*.

Tailleur, Modan, Chaymae Benaatia, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2025), « Diffusion-based spectral super-resolution of third octave acoustic sensor data: is privacy at risk? », *in*: *European Signal Processing Conference (EUSIPCO)*.

Tailleur, Modan, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2023), « Spectral trancoder: using pretrained urban sound classifiers on undersampled spectral representations », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Tailleur, Modan, Mathieu Lagrange, Pierre Aumond, and Vincent Tourre (2025a), « Audio Skim Generation for Environmental Audio Recordings », *in*: *(to be submitted)*.

— (2025b), *Enforcing Speech Content Privacy in Environmental Sound Recordings using Segment-wise Waveform Reversal*, DOI: 10.48550/arXiv.2507.08412.

Tailleur, Modan, Vincent Lostanlen, Jean-Philippe Rivière, and Pierre Aumond (2024), « Machine listening in a neonatal intensive care unit », *in*: *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Tailleur, Modan, Vincent Tourre, Pierre Amond, Mathieu Lagrange, and Ophélie Moreau (2025), « SoundSourceViz: A Visual Analytics Tool for Soundscape-Oriented Urban Planning with Acoustic Sensor Networks », *in*: *(submitted to Computer Graphics Forum)*.

Tan, Mingxing and Quoc Le (2021), « Efficientnetv2: Smaller models and faster training », *in*: *International Conference on Machine Learning (ICML)*.

Thakur, Sidharth and Andrew J. Hanson (2010), « A 3D visualization of multiple time series on maps », *in*: *International Conference on Information Visualisation (IV)*.

Thomas, Sinnu Susan, Sumana Gupta, and Venkatesh K. Subramanian (2017), « Smart surveillance based on video summarization », *in*: *IEEE Region 10 Symposium*.

Torija, Antonio J., Diego P. Ruiz, and A. F. Ramos-Ridao (2013), « Application of a methodology for categorizing and differentiating urban soundscapes using acoustical descriptors and semantic-differential attributes », *in*: *The Journal of the Acoustical Society of America (JASA)*.

Vartakavi, Aneesh, Amanmeet Garg, and Zafar Rafii (2021), « Audio summarization for podcasts », *in*: *European Signal Processing Conference (EUSIPCO)*.

Ville de Lorient (2013), *Plan de Prévention du Bruit dans l'Environnement (PPBE) – Ville de Lorient*, tech. rep., Ville de Lorient.

Wang, Yang (2019), *Deck.gl: Large-scale Web-based Visual Analytics Made Easy*, DOI: 10.48550/arXiv.1910.08865.

Wei, Penghui, Jiahao Zhao, and Wenji Mao (2021), « A graph-to-sequence learning framework for summarizing opinionated texts », *in*: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*.

Wilson, Denise L. and Robyn Crawford (1989), « Color Contrast Requirements for Legibility of Color Symbology Displayed against Color Backgrounds », *in*: *Proceedings of the Human Factors Society Annual Meeting*.

World Health Organization (2011), *Burden of disease from environmental noise: Quantification of healthy life years lost in Europe*, WHO.

Wu, Yusong, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov (2023), « Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation », *in*: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Wunderlich, Filipa Matos (2023), *Temporal Urban Design: Temporality, Rhythm and Place*, Taylor & Francis.

Xu, Changsheng, Xi Shao, Namunu Chinthaka Maddage, Mohan S. Kankanhalli, and Qi Tian (2004), « Automatically summarize musical audio using adaptive clustering », *in*: *IEEE International Conference on Multimedia & Expo (ICME)*.

Xu, Weihan, Paul Pu Liang, Haven Kim, Julian McAuley, Taylor Berg-Kirkpatrick, and Hao-Wen Dong (2024), *TeaserGen: Generating Teasers for Long Documentaries*, DOI: 10.48550/arXiv.2410.05586.

Xu, Xinyan, Tin Oberman, Andrew Mitchell, S. Huang, F. Aletta, and J. Kang (2023), « Effects of Visual Stimulus on Soundscape Perception: A Pilot Study Using Immersive Virtual Reality (IVR) », *in*: *Forum Acusticum*.

Yong Jeon, Jin, Pyoung Jik Lee, Joo Young Hong, and Densil Cabrera (2011), « Non-auditory factors affecting urban soundscape evaluation », *in*: *The Journal of the Acoustical Society of America (JASA)*.

Yuan, Li, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng (2019), « Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization », *in*: *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yue, Xiaoqi, Dan Feng, Desheng Sun, Chao Liu, Hongxing Qin, and Haibo Hu (2024), « AirPollutionViz: visual analytics for understanding the spatio-temporal evolution of air pollution », *in*: *Journal of Visualization*.

Zang, Sha-Sha, Hui Yu, Yan Song, and Ru Zeng (2023), « Unsupervised video summarization using deep non-local video summarization networks », *in*: *Neurocomputing*.

Zhang, Shu, Yingying Zhu, and Amit K. Roy-Chowdhury (2016), « Context-aware surveillance video summarization », *in*: *IEEE Transactions on Image Processing*.

Zhong, Bingzhi, Hui Xie, Tian Gao, Ling Qiu, Heng Li, and Zhengkai Zhang (2022), « The effects of spatial characteristics and visual and smell environments on the soundscape of waterfront space in mountainous cities », *in*: *Forests*.

Zhu, Ruirui, Liyun Yuan, Yanting Pan, Yefan Wang, Danping Xiu, and Weiyi Liu (2024), « Effects of natural sound exposure on health recovery: A systematic review and meta-analysis », *in*: *Science of the Total Environment*.

Zwicker, Eberhard and Hugo Fastl (2013), *Psychoacoustics: Facts and models*, vol. 22, Springer Science & Business Media.

COLLEGE ED SCIENCES
DOCTORAL DE L'INGENIERIE
PAYS DE LA LOIRE ET DES SYSTEMES

CENTRALE NANTES

Titre : Analyse à grande échelle des sources acoustiques, basée sur les données, pour l'analyse, la visualisation et le résumé audio de paysages sonores urbains

Mot clés : Réseau de capteurs acoustiques, Apprentissage profond, Classification de sources, Aménagement urbain, Analyse visuelle, Cube espace-temps, Résumé audio

Résumé : Les paysages sonores urbains jouent un rôle essentiel dans la qualité de vie en ville. Si le bruit peut avoir des effets néfastes sur la santé, des environnements sonores sains peuvent au contraire favoriser le bien-être, encourager les interactions sociales et contribuer positivement à l'expérience urbaine. Les réseaux de capteurs acoustiques, capables de collecter des données à grande échelle et sur le long terme, offrent de nouvelles opportunités pour analyser et représenter les environnements sonores urbains. Plusieurs réseaux ont déjà été déployés avec succès dans le cadre de projets de recherche, tels que SONYC aux États-Unis, Stadtlärm en Allemagne, ou encore CENSE à Lorient en France. Toutefois, traiter les données issues de ces réseaux demeure un défi. Ce document présente trois contributions visant à relever ce défi : (1) l'évaluation automatique de la présence de sources sonores à partir d'enregistrements spectraux compacts collectés par ces réseaux, (2) le développement d'un outil de visualisation interactive pour analyser des données acoustiques multi-sources à des fins de planification urbaine, et (3) la génération automatique de résumés audio courts à partir d'enregistrements environnementaux de longue durée. Les approches proposées combinent apprentissage automatique, acoustique environnementale et techniques de visualisation de données, avec pour objectif de permettre l'exploration des données audio environnementales tout en respectant les contraintes de compacité des données et de respect de la vie privée.

Title: Large-Scale Monitoring of Acoustic Sources Based on Data-Driven Approaches for Urban Soundscape Analysis, Visualization and Summarization

Keywords: Acoustic sensor network, Deep learning, Source Classification, Urban Planning, Visual Analytics, Space-Time Cube, Audio Summarization

Abstract: Urban soundscapes play a crucial role in shaping the quality of life in cities. While harmful noise can lead to adverse health effects, healthy sound environments can support well-being, encourage social interaction, and contribute positively to the urban experience. Acoustic sensor networks, capable of collecting large-scale and long-term data, offer new opportunities to analyze and represent urban sound environments. While several networks have been successfully deployed throughout the world as part of research projects such as SONYC in the US, Stadtlärm in Germany, and CENSE in Lorient, France, the meaningful assimilation of the data outputted by those networks remains a challenge. This document describes three major contributions that aim at addressing this issue: (1) the automatic assessment of sound source presence from compact spectral acoustic recordings of such sensor networks, (2) the development of an interactive visualization tool for analyzing multi-source acoustic data to support urban planning, and (3) the automatic generation of short audio summaries from long-form environmental recordings. The proposed approaches combine machine learning, environmental acoustics, and data visualization techniques with the aim of enabling the exploration of environmental audio data, while respecting constraints of data frugality and citizen's speech content privacy.