ReSilence

Retune the Soundscape of future cities through art and science collaboration

HORIZON- 101070278

D3.2

Audio, Visual and Multimodal Analysis Tools v2

Dissemination level:	Public
Contractual date of delivery:	Month 22, 30 June 2024
Actual date of delivery:	Month 26, 25 October 2024
Workpackage:	WP3: AI-based interactive technologies
Task:	T3.1: Multimodal movement analysis T3.1.1: Automated analysis of full-body individual movement expressive and emotional qualities T3.1.2: Automated analysis of full-body group movement expressive and emotion qualities T3.2: Sound and space sensing and Soundscape analysis T3.3: AI-based soundscapes
Type:	Demonstrator
Approval Status:	Final Draft
Version:	**1.0**
Number of pages:	62
Filename:	d3.2_resilence_AudioVisualandMultimodalAnalysisTools_v2.docx

Abstract

This deliverable reports the advanced techniques used in ReSilence for analysing all the acquired data: advanced techniques for the automated analysis of full-body individual and small group movement, soundscape modelling and analysis, machine learning algorithms and computer vision techniques that generate content from sounds and audio from visuals, using multimodal features.

European Commission

funded by the European Union

History

0.1	30/08/2024	ToC and first set of inputs	CERTH
0.2	15/09/2024	Inputs from partners	All involved partners
0.9	02/10/2024	Submitted for review	Alexander Shvets
0.10	16/10/2024	Final version after review	All involved Partners
1.0	17/10/2024	Quality control and submission	Nefeli Georgakopoulou

Author list

Organisation	Name	Contact Information
UNIGE	Antonio Camurri	antonio.camurri@unige.it
CERTH	Sotiris Diplaris	diplaris@iti.gr
CERTH	Nefeli Georgakopoulou	nefeli.valeria@iti.gr
CERTH	Paraskevi Kritopoulou	pakrito@iti.gr
CERTH	Eleftheria Lagiokapa	elagio@iti.gr

Executive Summary

This deliverable reports the advanced techniques used in ReSilence for analysing all the acquired data: advanced techniques for the automated analysis of full-body individual and small group movement, soundscape modelling and analysis, machine learning algorithms and computer vision techniques that generate content from sounds and audio from visuals, using multimodal features. Moreover, it reports the equipment designed to accommodate the needs of the artistic projects for collecting data.

Abbreviations and Acronyms

ADE	Art-Driven Experiment
AER	Audio Emotion Recognition
AI	Artificial intelligence
CDA	Chaussée d'Anvers
CDCML	Cross-Modal Deep Continuous Metric Learning
cGAN	conditional Generative Adversarial Network
CMRA	Cross-Modal Ranking Analysis
CNN	Convolutional Neural Network
CPU	Central Processing Unit
EMID	Emotionally paired Music and Image Dataset
FPGA	Field Programmable Gate Arrays
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Models
GPIO	General-Purpose Input/Output
GPU	Graphics Processing Unit
3D	three-Dimensional
2D	two-Dimensional
hfn	High frequency noise
HLUR	High level user requirements
HSV	Hue, Saturation, Value
IIR	Infinite ImpulseResponse
IMU	Inertial Measurement Unit
ITU	International Telecommunication Union
LAeq	Equivalent Continuous Sound Pressure Level
LDM	Latent Diffusion Model
lfn	Low frequency noise
LSTM	Long Short-Term Memory
MFCC	Mel-frequency cepstral coefficients
OS	Operation Systems
RF	Radio Frequencies
RFD	Radio-Frequency Interference
RNN	Recurrent Neural Networks
RTL-SDR	Realtek - Software Defined Radio
SBC	Single-Board Computers
SDR	Software Defined Radio
USB	Universal Serial Bus
USS	Universal Source Separation
VR	Virtual Reality
VSWR	Voltage Standing Wave Ratio
WLAN	Wireless Local Area Networks

Table of Contents

1. INTRODUCTION

One of the objectives of the ReSilence project is to involve and collaborate with artists in order to leverage multiple sources of inspiration, interdisciplinary collaboration, and build trust around AI & XR technologies. ReSilence supports Art-Driven Experiments (ADE) through Open Calls to artists, and the selected artists in ReSilence have access to AI and XR technology to reflect on novel uses and their impact on society. Furthermore, collaboration of S&T with the selected artists also helps in ensuring that the development process and system behaviour of the technologies explicitly acknowledge human values and needs, within the scope and objectives of the ReSilence project.

In this context, D3.2 focuses on the tools for audio, visual, and multimodal analysis provided by ReSilence partners to the selected artists. In particular, this deliverable presents an in-depth overview of the progress, tools, and equipment designed and developed for the Call 1 artistic projects. It also presents in detail the current objectives of the Call 2 artistic projects, outlining their development stages and the technologies being considered and tested. All of this highlights the achievements of WP3.

2. METHODOLOGY

This deliverable is part of WP3, focusing on the development and refinement of algorithms and techniques for real-time interactions in the immersive project use cases.

In particular, the following methodological directions characterise the work:

- Multimodal movement analysis, in individual users (T3.1.1) and in small groups (T3.1.2). The focus here is on the automated analysis of movement qualities, i.e., non-verbal full-body expressive, emotional and social signals;
- Methodologies and techniques for sound and space sensing and soundscape analysis and synthesis, with a particular focus on modelling perceived “annoyance” and “intrusiveness” of soundscapes, typical of city soundscapes (T3.2);
- Cross modal learning aiming at the development of an audio-to-image and image-to-audio synthesis model (T3.3);

3. RELATION TO USER REQUIREMENTS

The following table (Table 1) accumulates the user requirements that have been developed until now, based on the artists' expression of their needs. The user requirements are taken from the updated use cases section of D6.2 and are grouped under High level user requirements (HLURs).

Final HLUR	Final HLUR Title	Final HLUR Description
HLUR 1	Processing of audio files	Artists can isolate sounds of their choice in multiple frequencies, as well as analyse specific sound qualities
HLUR 2	Real time data analysis and feedback	Artists can use real time data analysis feedback to directly adapt and assess their prototypes.
HLUR 3	Multiple data and signal collection	Artists can record, track and measure physiological data as well as signals of movement
HLUR 4	Multiple data and signal analysis	Artists can analyse data from online sources, physiological data as well as signals of movement
HLUR 5	Data synchronisation	Artists can utilise synchronised sources of data inputs and outputs
HLUR 6	Data translation/visualisation	Artists can externalise/visualise sonic and physiological data
HLUR 7	Artistic installation user feedback	Artists can collect and analyse user feedback after they experience their installation
HLUR 8	Audio recording quality	Artists can have audio files of high quality and of at least 2-3 minutes duration without disruptions
HLUR 9	Wearable sensor positioning adaptation	Artists can adapt the positioning and quantity of wearable sensors to allow as free movement to the end user as possible
HLUR 10	Aesthetic evaluation of sound and experience	Artists can use scientific methods to explain the psychological, neuronal and socio-cultural basis of aesthetic perceptions of sound and music
HLUR 11	Use generative sound	Artists can use generative AI models to create

	models	music, voice clones and musical compositions
HLUR 12	Geolocate and track objects and sound sources	Artists can be able to geolocate humans, objects and sound sources through APIs and apps
HLUR 13	Sound navigation system for audience including visually impaired	Artists can use sound navigation systems to guide and audience in groups or assist guidance of the visually impaired

Table 1: Analysis of the High-level user requirements

4. MULTIMODAL MOVEMENT ANALYSIS

In this section is summarised the achievements on ReSilence's T3.1, regarding supported provided to the artists on projects that require automated analysis of full-body individual movement, expressive, and emotional qualities.

4.1. Related work

The project starts from the EyesWeb movement analysis software library, which provides a series of software modules integrated in the EyesWeb platform for the analysis of mid-level movement qualities, including individual movement qualities (e.g., Origin of Movement OoM (Kolykhalova et al., 2020), Fragility, Lightness (Niewiadomski et al., 2019) as well as small groups movement qualities, with a particular focus on synchronisation and entrainment (e.g. (Sabharwal et al., 2022; Alborno et al., 2019). An overview of the conceptual framework for the software library is shown in the following picture.

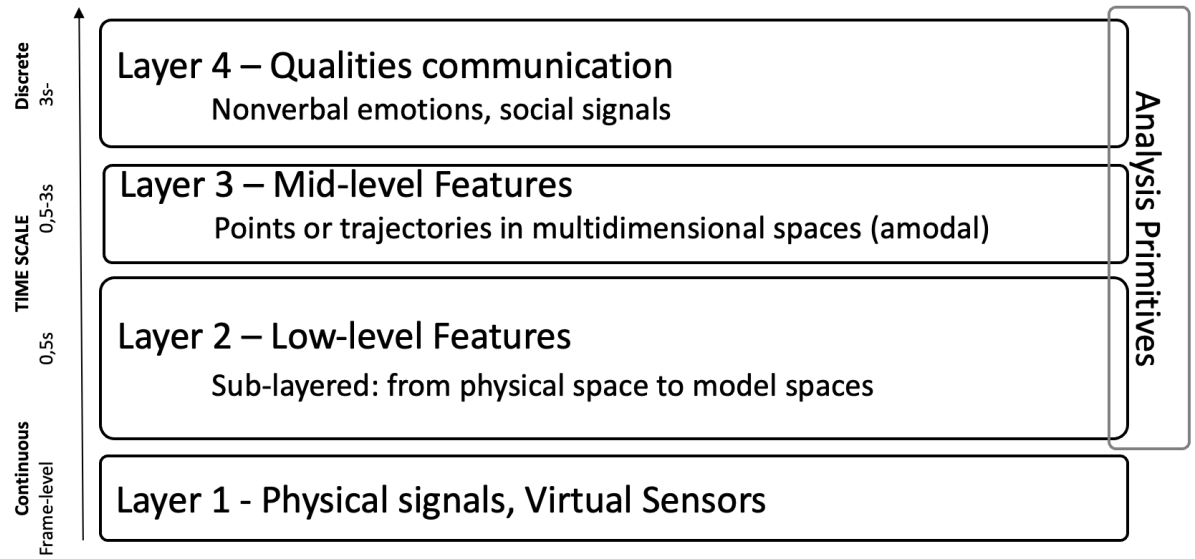


Figure 1: An overview of the conceptual framework for the software library

4.2. Hardware equipment

We developed a scalable movement analysis system, consisting of two setups: (i) a low-cost platform, based on depth-cameras (e.g., Kinect) and wearable sensors (IMUs); (ii) a lab setup, based on a 19-camera Qualisys motion capture system. The latter is adopted for lab experiments at the Casa Paganini research lab at UNIGE. The former is used for experiments in-the-wild, e.g., the installations of the project of the composer Andrea Cera. The setup at our research centre has been presented and discussed with various artists participating in ReSilence, with a special focus on the interaction design and the choice and refinement of the suitable sensing technologies and interaction paradigms in their projects.

4.2.1. Hardware equipment to support art driven experiments of other artistic projects

We had an intense collaboration and participation to the interaction design development process with Andrea Cera, and also a series of meetings with Brigitta Muntendorf (physical meeting in Venice, with Beatrice de Gelder, in occasion of her concert at Biennale of Venice),

Lea Luka Sikau (contribution to the interaction design of her interactive installation), and Loukia Tsafulia and Alfonso Severino, which included a meeting in presence on 25 June 2024 at our research centre InfoMus-Casa Paganini, see pictures below.

(a)

(b)

Figure 2: Collaborations with artists (on stage) during interaction design development

The premise of the InfoMus-Casa Paganini of UNIGE, used in the collaboration with several artists (Andrea Cera, Lea Luka Sikau, Loukia Tsafulia and Alfonso Severino, and Paul Louis) in both the interaction design of their artistic projects and in related scientific experiments with Andrea Cera.

4.3. **Dataset creation**

A number of voluntary dyad participants were recruited and recorded using a 19-camera Qualisys mocap setup at Casa Paganini/UNIGE. We recorded a series of movement recordings of about 10 dyads performing joint emotional tasks, in the premise of InfoMus-Casa Paganini at UNIGE (see picture above), using the Qualisys mocap setup. The goal was to obtain a dataset to validate movement analysis techniques adopted in the project of the composer Andrea Cera. A specific research challenge is the following:

- does an intrusive soundscape cause a degradation/decrease of the quality of individual as well as group movement qualities in human participants performing an emotional task?

In collaboration with the composer Andrea Cera, we recorded a dataset of soundscapes characterised by high intrusiveness, neutral, and calm naturalistic (non-intrusive) soundscapes. The techniques for the design of such soundscape recordings emerged from the collaboration between UNIGE and Andrea Cera. This has been used as sound background stimuli to the dyads in the joint emotional movement recordings of participants in the scientific experiment.

4.3.1 **ANDREA CERA Sound DataSet**

In this section we give a short description of the sound dataset created by Adrea Cera used in both scientific experiments and in his artistic project.

1. a collection of 36 [high intrusiveness] + 36 [low intrusiveness] individual sounds

These sounds were used by CERTH for the generation of image database, and by InfoMus-CasaPaganini/UNIGE in a montage described below.

- the 36 [high intrusiveness] sounds are short recordings made by Andrea Cera with a Zoom H2 during the last 10 years, and a few simulations. These sounds include:

- cars, motorbikes, trucks passing by
- jackhammers, mechanical arms, other construction sounds
- cooling units, refrigerators, other ventilation sounds
- skateboards, containers rolling, hammering
- simulation of smartphone sounds and warning beeps
- simulation of music being played inside closed spaces (club, store, etc.)

Each sound is accompanied by a 4-item list describing its spectral centroid, spectral sharpness, spectral skewness and roughness average + a categorisation following a taxonomy of urban sounds.

All sounds are normalised to peak in a range between -1.2 dB and -0.3 dB.

- the 36 [low intrusiveness] sounds are obtained by applying a series of transformations to the high intrusiveness sounds (filterings, changes in temporal shape, addition of components, etc.), aimed at lowering their intrusiveness. As for their counterpart, these sounds too are accompanied by 4 item lists describing spectral centroid, spectral sharpness, spectral skewness and roughness averages.

The peak amplitude of these sounds varies inside a range of 7 dB.

Since intrusiveness is not yet a universally established parameter, and the operations of transformation are quite radical and experimental, the sound designer (Andrea Cera) decided to opt for a level of tolerance in the setting of the overall intensity of the transformed sounds.

2. a 20'' long montage from the 36 [high intrusiveness] sounds

These sounds have been designed to be used as soundscapes for the participants in the scientific experiment developed by InfoMus-CasaPaganini/UNIGE (dyads performing joint emotional actions).

This montage consists of 21 [high intrusiveness] sounds. The materials are organised in the following way:

- a principal foreground with a succession of 13 sounds of passing cars, motorbikes, trucks. All sounds of this layer are played back at the original level, except for 2 of them, played back 1 dB softer to avoid excessive pre-clipping caused by superposition of fade in / and fade out.
- secondary layer with 8 other sounds used to add contextual elements (construction sounds, klaxon, etc.).

A very light limiter was placed in the master bus to prevent other clipping.

In the InfoMus-CasaPaganini/UNIGE experiment, measured with a class B sound metre, in the point around which the participants moved, this soundscape reached occasional peaks at 80 dBA, and in average stayed around 75 dbA. The calibration sound (sound n.04) peaked at 80 dbA.

This sound file is accompanied by 4 .csv / txt files containing an analysis of spectral centroid, spectral sharpness, spectral skewness and roughness. Combined with the motion capture files from the experiment, these .csv files represent a useful material to investigate is intrusive sound qualities have or not an effect in movement qualities.

3. a 20'' long montage low intrusiveness sounds

This is derived from InfoMus-CasaPaganini/UNIGE project *DanzArTe*, also used in the ReSilence scientific experiment in CasaPaganini.

This is a rendering of the same DanzArTe session used to teach the movement to the subjects in CasaPaganini's experiment (NB: in the training session there is no sound).

The sounds in DanzArTe were designed following low intrusiveness guidelines, regarding timbral and dynamical dimensions, and also with the intention of nudging the participants towards an ideal cadence and intention of movement.[1]

Measured with a class B sound metre, in the point around which the participants move, this soundscape stayed around 50/55 dbA, with one occasional peak at 60 dB.

As for the high intrusiveness sounds, this rendering too is accompanied by 4 .csv / txt files containing an analysis of spectral centroid, spectral sharpness, spectral skewness and roughness. Combined with the corresponding motion capture files from the experiment, these .csv / txt files allow to compare the effect of high/low intrusiveness sounds in movement qualities.

All the analyses were performed with the library libxtract + Sound Visualiser (Jamie Bullock / Queen Mary University London) and MaxMSP library ZSA descriptors (Mikhail Malt and Emmanuel Jourdan / IRCAM.

4.4 **Future considerations**

The work with artistic projects is a source of inspiration for novel developments and refinements of computational models of analysis of movement, at both individual and group level. In this direction we are working on the analysis of the recorded movement data, with the aim of evaluating and validating refined versions of the software modules for the analysis of movement, using as testbed the artistic project of Andrea Cera.

According to the main research objectives, the recorded dataset will be used to measure if/how intrusive soundscapes influence/perturbate joint emotional tasks.

[1] More details on the DanzArTe project at https://www.youtube.com/watch?v=OxJ0e8zums8

5. SOUND AND SPACE SENSING AND SOUNDSCAPE ANALYSIS

In this section we summarise our analysis of the state-of-the-art technologies and tools that could be potential candidates for the projects of Call 2 artists (discussed in Section 5.1) or were utilised for the automated analysis of sonic features in soundscapes for Call 1 artists projects (in sections 5.2 and 5.3). This includes both third-party and partner-developed software or software libraries. Additionally, equipment specifically designed or adapted for sound and space sensing, as well as soundscape analysis, is described in detail in Section 5.2.

A detail description for each artist involved in the collaboration with the artistic project and related scientific experiments of WP3, is provided, and more specifically, for Guillem Serrahima (Call 2), Caroline Clauss (Call 1) and Andrea Cera (Call 1).

5.1 Related work to support the art driven experiments of 2nd call artists

This subsection describes the approach that is currently being designed and developed in CERTH, to use in Guillem Serrahima's artistic project (Call 2), that requires space sensing and recording, and radioscape analysis.

5.1.1 Supporting Guillem Serrahima's experiments and project challenges

At this stage, Guillem Serrahima gravitates towards exploring an aspect of soundscape analysis that focuses on higher frequencies than the audible sound, and one of the directions he is heading to is mapping waves in the scale of radio frequencies (RF). In general, *Radio art* refers to works created by artists using the medium of radio, and it is applied to sound-based creations incorporating elements of radiophonic language—such as voice, words, music, sound effects, and silence—to craft aesthetic messages and evoke emotional responses from listeners during a broadcast (Soto-Sanfiel et al.,2022). These works are produced and broadcasted by utilising the radio infrastructure and technologies. In contrast to this approach, and remaining in the framework of the ReSilence aims, Serrahima envisions exploring a rather not so common concept, the "*Radioscape"*.

By the definition given by the sound artist Edwin van der Heide[2], Radioscape is "an immersive environment that redefines the radio medium, establishing a new bodily relationship to the medium and adding a new layer to a region of a city". Taking inspiration from this work, this analysis will focus on the existing radio waves that are present in a natural space. These radio waves may be generated either by human activities and technology like telecommunication devices (e.g., mobile phones, satellite signal), media transmissions (e.g., radio, television, etc), and wireless network technology (Wireless Local Area Networks (WLANs), including Wi-Fi, Bluetooth, LoRa, and Z-Wave). Alternatively, radio waves can occur naturally originating from cosmic sources such as celestial bodies (e.g., planets, quasars, galaxies, etc) that emit signals as part of cosmic radiation.

To achieve the recording and capture of Radioscape, appropriate equipment is necessary to be designed or developed for radio signal capture and modulation according to the needs of the project. A RF detector (or else RF power detector, or RF responding detector) is needed to detect the presence of RF waves. Signals collected from radio telescopes and provided by

[2] https://www.evdh.net/radioscape/

observatories may be also utilised for the needs of the project.

5.1.1.1 Related work

There have been several attempts to record the radio waves and modulate them to acoustic sounds in various artistic projects, that utilise specific hardware parts. Usually, these projects require an antenna as a receiver or a transmitter to produce the artistic result. In more detail, special music instruments like the ***Theremin*** (Glinsky, 2000) have been developed to modulate the RF oscillators' signals using the performer's hand movements (in practice the human body capacitance) to control pitch and volume, thus transforming the wave into audible sound. Experimenting with different radio parts, ***Feedback loops*** involve the recirculation of audio signals through a system of transmitters and receivers, allowing complex sound modulation (Burtner, 2003). This method can create layers of sound that evolve and transform in response to environmental and performer interactions.

In a different approach that also involves visual feedback, the ***mini-FM*** device was designed to utilise FM transmitters to produce intricate audio-visual effects (Kogawa, 2008; Hall, 2018). These effects include dynamic soundscapes and evolving visual patterns, which are influenced by interference and other modulating factors, creating an engaging and multi-sensory experience for the audience. Likewise*,* ***Environmental interaction*** uses an array of FM transmitters and receivers to allow the audience to influence the audio-visual environment (Friz, 2011; Friz, 2011). By moving within the space, participants can modulate and transform the sound and visual output, creating an immersive and participatory experience.

Other artists are developing intriguing hybrid works that integrate radio instruments with various sound-making practices, while maintaining an emphasis on feedback within a circuit of bodies and devices. There are also many artistic installations and performances that use a similar technological manipulation to produce unconventional audio effects like the ***Circuit bending***, but these approaches do not focus on Radio waves. Notably the sound artist Edwin Van der Heide in collaboration with Paul Mourus and Alexei Blinov, applied the Mid-Side stereo recording technique to a set of antenas + receivers to be able to listen to electromagnetic signals in "stereo" (van der Heide, 2000).

Regarding the “Radioscape” project CERTH is speculating with the possibility of a similar approach for another spatial audio rendering format which is Ambisonics. To capture the electromagnetic signal, an economical approach is being considered, that revolves around special devices called Software Defined Radio.

Software-defined radio (SDR)

Traditionally, radio components like modulators, demodulators, and tuners have been implemented using analogue hardware. However, with the advent of modern computing and analogue-to-digital converters, many of these components can now be implemented in software, leading to the concept of software-defined radio. This shift enables more accessible signal processing, allowing for the production of cost-effective wide-band scanner radios[3].

Most affordable SDR receivers today are based on the Realtek RTL2832U[4] chipset, originally designed for receiving digital television, yet there are more options available on the tuner (e.g.

[3] https://www.rtl-sdr.com/about-rtl-sdr/
[4] RTL2832U Datasheets: https://homepages.uni-regensburg.de/~erc24492/SDR/RTL2832U.pdf, https://homepages.uni-regensburg.de/~erc24492/SDR/Data_rtl2832u.pdf

R820T2, R828D, E4000, or the less frequently used chips, FC0013, FC2580 MSI2500/MS101, etc). These small devices serve as the foundation for nearly all low-cost SDR units available today. The chip is mounted alongside all the necessary supporting electronics on a circuit board that's no larger than a USB thumb drive. This compact package includes a USB port, which supplies all the required power and enables direct communication with the computer it is connected to.

Field Programmable Gate Array (FPGA)-designed as SDR systems have also been attempted (Panda et al., 2014) in an effort to manage the complexity of modern SDR systems, yet with some drawbacks like power dissipation or communication delay (Maheshwarappa et al., 2017). On the other hand, projects aiming at improved performance are utilising FPGAs as a means to reduce the communication volume between SDRs and hosts (Rahman & Islam, 2016). Designing FPGAS as SDR's systems usually takes place

5.1.1.2 Hardware equipment

RTL-SDR (Realtek - Software Defined Radio) is a very popular product that the artist seems to favour in the current phase of the project (Figure 3). No individual or company has full ownership of RTL-SDR and all the software and hardware that support it. However, the discovery that certain TV dongles could be used for software-defined radio (SDR) was the result of combined efforts[5] by Antti Palosaari[6], Eric Fry, and Osmocom[7], with significant contributions from Steve Markgraf[8]. Osmocom developed the first RTL-SDR driver, which was released as open source.

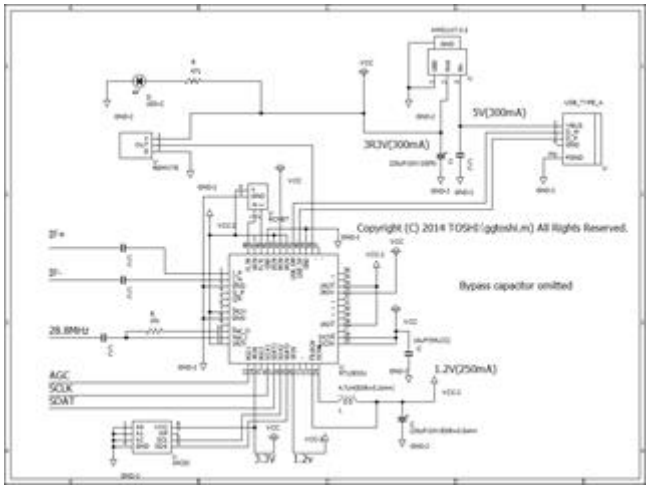


Figure 3: Reverse engineered RTL-SDR schematic created by GGToshi[9] (Copyright© 2014 TOSHI (ggtoshi.m) All Rights Reserved)

DVB-T dongles utilising the Realtek RTL2832U chipset can be repurposed as affordable SDR devices (an Eric Fry discovery) since the chip enables the transfer of raw I/Q samples to the host, originally intended for DAB/DAB+/FM demodulation. In general, most GUI-based software-defined radio programs will require at least a dual-core processor. However,

[5] https://charleslabs.fr/en/project-Getting+started+with+RTL-SDR
[6] https://palosaari.fi/linux/
[7] https://github.com/osmocom/rtl-sdr
[8] https://wiki.steve-m.de/, https://osmocom.org/projects/rtl-sdr/wiki/Rtl-sdr
[9] http://ggtoshi.at.webry.info/201406/article_6.html

command-line tools and ADS-B decoders may function on less powerful hardware. Single-board computers (SBC) like the Raspberry Pi 3, as well as Android mobile devices, can also run several of these applications.

A wide range of software is available for the RTL2832. Most user-level applications depend on the librtlsdr library, which is included in the rtl-sdr codebase. This codebase not only provides the library but also includes several command-line tools such as rtl_test, rtl_sdr, rtl_tcp, and rtl_fm. These tools utilise the library to detect RTL2832 devices and perform basic data transfer operations to and from the device.

5.1.1.3 Future considerations

When selecting an antenna for a particular application, there are several key features to consider. These factors influence the antenna's performance and suitability for specific environments or devices. The receiver's characteristics will be mainly defined by the signal's wavelength. Therefore, the project's final focus will define which/what Antennas will be used. The generation and transmission of radio frequency bands are strictly regulated by national laws and coordinated internationally by the International Telecommunication Union[10] (ITU), an agency of the United Nations based in Geneva. The ITU manages the shared global use of the radio spectrum, promotes cooperation in assigning satellite orbits, improves telecommunication infrastructure in developing countries, and assists in developing global technical standards. This organisation plays a critical role in preventing interference between different users by coordinating the allocation of radio frequencies.

In more detail, the ITU allocates different sections of the radio spectrum, called radio frequency (RF) bands, for various communication technologies and services, ensuring non-overlapping usage for applications like broadcasting, mobile radio, or navigation. A radio frequency band is a specific, continuous section of the radio spectrum where channels are typically designated for specific uses.

The organisation has defined around 40 radio communication services in its Radio Regulations and has a band plan for each frequency range to maintain compatibility between transmitters and receivers. Additionally, the ITU divides the radio spectrum into 12 bands based on wavelength and frequency, with specific naming conventions for each. Table 2 summarises the existing bandwidths defined by ITU:

Band name	Frequency	Wavelength	Example Uses
Extremely low frequency (ELF)	3–30 Hz	100,000–10,000 km	Communication with submarines
Super low frequency (SLF)	30–300 Hz	10,000–1,000 km	Communication with submarines
Ultra-low frequency (ULF)	300–3,000 Hz	1,000–100 km	Submarine communication, communication within mines

[10] https://www.itu.int/en/Pages/default.aspx

Very low frequency (VLF)	3–30 kHz	100–10 km	Navigation, time signals, submarine communication, wireless heart rate monitors, geophysics
Low frequency (LF)	30–300 kHz	10–1 km	Navigation, time signals, AM longwave broadcasting (Europe and parts of Asia), RFID, amateur radio
Medium frequency (MF)	300–3,000 kHz	1,000–100 m	AM (medium wave) broadcasts, amateur radio, avalanche beacons
High frequency (HF)	3–30 MHz	100–10 m	Shortwave broadcasts, citizens band radio, amateur radio and over-the-horizon aviation communications, RFID, over-the-horizon radar, automatic link establishment (ALE) / near-vertical incidence skywave (NVIS) radio communications, marine and mobile radio telephony
Very high frequency (VHF)	30–300 MHz	10–1 m	FM, television broadcasts, line-of-sight ground-to-aircraft and aircraft-to-aircraft communications, land mobile and maritime mobile communications, amateur radio, weather radio
Ultra-high frequency (UHF)	300–3,000 MHz	1–0.1 m	Television broadcasts, microwave oven, microwave devices/communications, radio astronomy, mobile phones, wireless LAN, Bluetooth, ZigBee, GPS and two-way radios such as land mobile, FRS and GMRS radios, amateur radio, satellite radio, Remote control Systems, ADSB
Super high frequency (SHF)	3–30 GHz	100–10 mm	Radio astronomy, microwave devices/communications, wireless LAN, DSRC, most modern radars, communications satellites, cable and satellite television broadcasting, DBS, amateur radio, satellite radio
Extremely high frequency (EHF)	30–300 GHz	10–1 mm	Radio astronomy, high-frequency microwave radio relay, microwave remote sensing, amateur radio, directed-energy weapon, millimetre wave scanner, wireless LAN (802.11ad)
Terahertz or tremendously high frequency (THz or THF)	300–3,000 GHz	1–0.1 mm	Experimental medical imaging to replace X-rays, ultrafast molecular dynamics, condensed-matter physics, terahertz time-domain spectroscopy, terahertz computing/communications, remote sensing

Table 2: ITU defined radio spectrum bands

Besides the frequency range, other factors important to this project that need to be examined are the antenna's Gain and Beamwidth, which will define the focus of the area reception (a narrow area reception will help mapping the area in more detail). Discussions with the artist

will help in realising the priority of other factors like the Polarisation that controls the maximum signal strength and reception since the transmitter and receiver should have the same polarisation (mismatched polarisation results in signal loss), or the Voltage Standing Wave Ratio (VSWR) that indicates how well the antenna is matched to the transmission line or radio and of course its Efficiency. Gravity will be also given to Size and Form Factor since portability may pose an issue.

5.2 **Developing Caroline Clauss' "Audio Recording Toolbox"**

In this subsection is provided a description on the support provided by CERTH for Caroline Clauss' project regarding sound and space sensing and soundscape analysis.

The urban soundscape comprises both natural and human-made acoustic signals. In projects like Caroline Clauss', which involve human activities such as speech and anthropogenic noise, field recording is a sensitive matter due to GDPR regulations. Thus, the challenge lies in capturing human voice activity without removing it from the soundscape during recording and before storage. This becomes more complex given the extended duration of recording sessions and the associated power consumption issues.

Our proposed solution involved "blurring" voices to alter identity characteristics and make speech incomprehensible. Human voices typically span in frequencies between 20Hz and 20kHz, overlapping with other urban sounds. Therefore, voice detection, source separation, and manipulation are essential.

Recording in real-time while complying with GDPR necessitates specialised equipment. Implementing voice-blurring on a low-cost, low-energy, portable device presents additional challenges, as it must preserve ambient voices without merely cutting them off. The solution must offer real-time voice distortion during recording and detection to safeguard privacy while maintaining environmental sound data integrity.

5.2.1 **Related work**

In continuation to the state-of-the-art research conducted for the previous deliverable 3.1, supplementary material to support the needs of the project, have been studied, regarding human anonymisation.

V-CLOAK is a sophisticated real-time voice anonymisation system that balances the needs for intelligibility, naturalness, and anonymity. V-CLOAK's one-shot generative model modifies the input audio at multiple frequency levels, using a loss function which factors in anonymity, intelligibility, and psychoacoustic naturalness. This multi-level approach ensures that while the speaker's identity is anonymised, the voice retains its timbre and remains natural to listeners. V-CLOAK also provides flexibility, offering both untargeted and targeted anonymisation, allowing it to be applied in diverse settings like instant messaging or social media (Deng et al., 2023).

For voice anonymisation on devices with limited computational resources like a Raspberry Pi, simpler and more computationally efficient methods are necessary, as complex systems like V-CLOAK are not feasible. The approach detailed in Voice Anonymization in Urban Sound Recordings offers a practical solution by using a U-Net model for source separation followed by two anonymisation techniques: low-pass filtering and MFCC (mel-frequency cepstral coefficients) inversion. These techniques blur the speech content and identity while

preserving the acoustic scene, and their lightweight computational demands make them suitable for devices like the Raspberry Pi. By focusing on frequency manipulation and selective filtering, this method achieves real-time voice anonymisation without overburdening the device (Cohen-Hadria et al., 2019).

An additional challenge is maintaining the loudness and presence of the voice while anonymising its content. Adaptive Loudness Compensation focuses on adjusting audio signals based on changes in playback volume to preserve spectral balance across frequencies, ensuring that softer sounds retain their presence. The technique employs low-order infinite impulse response (IIR) digital filters to adapt audio based on human perception, compensating for reduced loudness without sacrificing audio quality. This method could be particularly useful in voice anonymisation systems, where maintaining the prominence of the voice in the overall mix is crucial, even if the intelligibility is compromised (Fierro et al., 2019).

(a) (b)

Figure 4: a) Raspberry Pi 4B connected with peripherals, using the AudioMoth as a recording device, b) Portable set up in a protective box, ready to be deployed to the urban environment

5.2.2 **Hardware equipment**

Within the ReSilence framework, specific requirements for recording devices have surfaced to align with a) the artists' creative vision, b) compliance with GDPR legislation, and/or c) utilisation of equipment already possessed by residency artists. More precisely, devices were combined in a portable and functional design that allowed data acquisition. The AudioMoth sensor (Hill et al., 2019) for audio recording, and the Raspberry Pi 4B[11] SBC, were integrated to achieve real-time sound analysis before storing audio data collected from urban environments.

Caroline Clauss selected to utilise power banks to power up her Raspberries. Support was provided on various products found/selected by the artist, such as novel products that use solar panels to charge complementary the power bank, while installed for recordings. An estimation of the power consumption of the Audio Recording Toolbox also took place. Power Banks of 20000mAh were selected to support scheduled recordings that run without supervision on a 24-hour basis.

In total, the artist purchased four devices, aiming to collect audio recordings that would be

[11] https://www.raspberrypi.com/products/raspberry-pi-4-model-b/specifications/

able to be mixed for creating an ambisonic environment. Ambisonic is a full-sphere sound format, or to explain more simply, a 360° surround soundscape (Gerzon, 1973). These recordings were mixed in an artistic synthesis that will be part of the artist's exhibition.

5.2.3 Software to support art driven experiments

As already mentioned in D3.1, the Raspberry Pi is a cost-effective single-board computer with various generations offering integrated CPU and GPU, on-board memory, camera port, and GPIO pins for connecting electronics. It functions like a standard computer, supporting various operating systems (Linux Desktop environment, Windows 10 IoT, and Android) and external devices like keyboards, mice, screens, and microphones. Additionally, it can autonomously run scripts in multiple programming languages.

In ReSilence, the Raspberry Pi 4B model with an 8GB RAM is used for audio signal recording and processing, paired with the AudioMoth microphone for audio input. The equipment was configured to support real-time GDPR-compliant sound analysis to ensure privacy in urban areas.

An OS that is well-optimised for the hardware can improve overall performance, stability, and responsiveness. This is particularly important for devices like the Raspberry Pi, which has relatively modest hardware specifications compared to full-sized PCs. Therefore, Raspberry Pi OS was selected to support the device. Therefore, the devices were flashed using the Raspberry Pi Imager that allows early/easy customisation of the Raspberry Pi OS[12] settings. Raspberry Pi OS, is a Linux distribution specifically designed for the Raspberry Pi platform. It is built on Debian Linux, one of the foundational versions of Linux, from which Ubuntu is also derived.

KDE (K Desktop Environment) is an example of a desktop environment. Unlike Windows or Mac, where the desktop environment is integrated into the operating system, Linux separates the desktop environment from the core operating system, allowing it to be installed on top of Linux. Most Linux distributions come with a default desktop environment. Originally, Raspbian (first versions of Raspberry Pi OS) used LXDE, a lightweight desktop environment suited for the Raspberry Pi's low-power hardware. More recent versions include the PIXEL desktop environment, although there is also a Lite version of Raspbian that comes without a desktop environment at all.

Additional Python3, packages and GUI were installed to support the algorithms that performed source separation (Universal Source Separation (USS) (Kong et al., 2020) and blurring. Below is presented a list of the python packages installed on the device (Table 3).

Operating System (OS)	Raspberry Pi OS
Programming language	Python 3
Blur packages	openpyxl
	pyloudnorm
Source separation packages	lightning 2.0.0
	h5py 3.8.0

[12] https://www.raspberrypi.com/documentation/computers/os.html

	librosa 0.10.0.post2
	pandas 1.5.3
	panns_inference 0.1.0
	tensorboard 2.12.2
	einops 0.6.1
Audio I/O packages	portaudio19-dev
	Pyaudio

Table 3: Table 3: Raspberry Pi System set-up

5.2.3.1. **Voice Blurring**

Various techniques were explored and tested for blurring the human voice. Two key factors were considered:

1. Altering the timbre of the voice, and
2. Ensuring that the intelligibility of the recorded speech samples was not preserved.

To achieve timbre masking, several filters were tested, including different low-pass filters that maintained the fundamental frequency band of the voice. There was also experimentation with an anonymisation method using mel-frequency cepstral coefficients (MFCC) within a low-pass filter, but it resulted in the loss of voice essence and introduced too much noise, rendering it ineffective for AI soundscape analysis. Additionally, loudness compensation was researched to maintain the same perceived sound presence before and after modifying parts of the signal, though this affected how the voice's loudness was perceived by the human ear.

In terms of reducing intelligibility, although the previous filtering methods were highly effective, further experiments were conducted by adjusting the batch size for sample shuffling. In the final implementation, the voice blurring integrated into the device uses a low-pass filter with loudness compensation.

5.2.3.2. **Source Separation**

USS algorithm is provided with pre-trained models that were effective for the needs of the human voice separation, therefore it was not necessary for further training. In the system, to start recording the Python script "audiomoth_rec.py" is called (Figure 5). This script was developed to communicate with the AudioMoth microphone to start the recordings and then call the USS algorithm to separate the human voice. The collected sound is analysed by the USS to detect human voices/speech and then processed to subtract the human voice signal. Consequently, the human voice signal is filtered appropriately to blur the signal, rendering the identity of the speaker/speakers or the content of the speech incomprehensible.

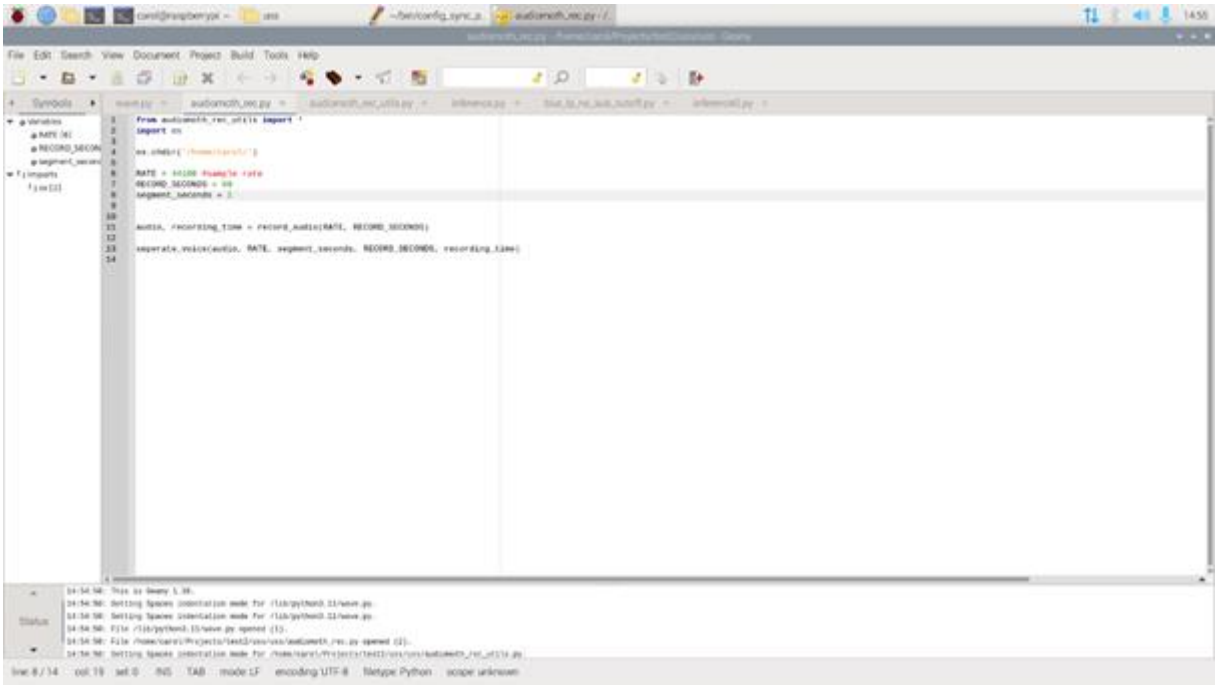


Figure 5: Editing “audiomoth_rec.py”

The scripts were adjusted to the needs of the project by creating a framework in which the end user could set the required values to crucial parameters like sample rate, recording duration, and the duration of the parts of the recordings that were analysed step by step. In more detail, to set the recordings, the end user needs to define in audiomoth_rec.py the values of the following parameters:

- RATE: Sets the Sample rate in Hz. Values above 44100 are not advisable since the higher the sample rate, the time needed to process the recorded sound increases.
- RECORD_SECONDS: Sets the duration of the audio file to be recorded in seconds. The longer the duration, the time needed to process the recorded sound increases.
- segment_seconds: Sets the separation window in length. It is practically the duration of a segment in seconds. The higher this value, the more accurate the voice recognition. Additionally, the lower its duration the longer the processing time.

The blurring filters whose performance was closer to what the artist described as optimal result, were selected and integrated in the USS algorithm, to ensure that the original human voice characteristics are never stored in the device’s microSD card.

5.2.3.3. **Ambisonic Recordings**

For ambisonic recordings, the four Raspberry Pi devices need to perform synchronised recordings. Therefore, various approaches were implemented on the system, to select the solution which provides the best performance outcome.

The initial approach was focused on uDev Rules, so that each time a usb device is connected – in this case the AudioMoth microphone – a script that controls and schedules the recordings according to the devices clock, would be activated. Unfortunately, the Raspberry Pi OS environment imposed certain limits. The USS algorithm needs several minutes to complete,

and its duration is strongly related to the sample rate, the content of the audio file that is inputted by the recording (are there human voices present, are the human voices discernible), the recording duration, etc. The script required more than 5 minutes to execute successfully – a very long duration – and this fact resulted in script termination by the system. Manipulating the timeout in the script did not resolve the issue, so different approaches were adopted.

Experimentation on a) Systemd Services and b) Cron Services took place. Cron scheduling provided adaptability to the end user, rendering the device more scalable, therefore it was selected as a final approach. Cron is a time-based job scheduler in Unix-like operating systems. It allows running scripts or commands at specified intervals. The Cron service runs in the background and checks scheduled tasks (jobs) to execute them at the appropriate times. Additionally, crontab (which stands for "cron table") is a file where you define the schedule for the cron jobs. Each user can have their own crontab file, and there is also a system-wide crontab. By editing the crontab file it is easy to specify when and what scripts or commands will run.

According to the schedule requested by the artist, the crontab prepared for the Audio Recording Toolbox had the following structure:

```
XDG_RUNTIME_DIR=/run/user/1000
SHELL=/bin/bash

# Morning recordings script
45 5 * * * sudo /sbin/reboot
2 6 * * * /home/carol/bin/morn-sync-rec-launcher.sh

# Noon recordings script
35 16 * * * sudo /sbin/reboot
49 16 * * * /home/carol/bin/noon-sync-rec-launcher.sh

# Night recordings script
10 23 * * * sudo /sbin/reboot
25 23 * * * /home/carol/bin/midnight-sync-rec-launcher.sh

# freeSync recordings script
#25 23 * * * /home/carol/bin/sync-rec-launcher.sh

# Shutdown
#57 4 * * * /usr/bin/bash /home/carol/bin/shutdown-raspi.sh
```

Each line in this crontab file represents a scheduled task that calls a specific script to schedule the recordings. Four scripts were developed to call the “audiomoth_rec.py” script and cover the user's needs.

1. “morn-sync-rec-launcher.sh” and “noon-sync-rec-launcher.sh”: to run during the day
2. “night-sync-rec-launcher.sh”: to run while crossing the midnight (starting before midnight, ending after midnight)
3. “sync-rec-launcher.sh”: to run continuously until and if the end user wishes to stop the recordings
4. “shutdown-raspi.sh”: to safely terminate the device automatically

Since the hour changes at midnight, it was difficult to create a universal script that would both run just during the interval hours [00:00 - 23:59], and in hours that cross the midnight. Also,

in case there is a connection to a stable and not portable power source, the end user may wish to set an open-end recording session that may cross the midnights if necessary or not. The optimal solution in this case for the end user requires not needing to handle too many parameters, thus the “sync-rec-launcher.sh” script was created.

All scripts are synchronised to run the Python script at specific intervals defined by the variable “recording_frequency”. This periodicity aims to ensure ambisonic recordings that render all devices synchronised to the second. Additionally, they include Error Handling that checks if the Python script “audiomoth_rec.py” itself is encountering any errors that might cause it to terminate prematurely. They were developed so that error handling mechanisms are in place to capture and log any errors appropriately. Finally, logging is implemented throughout the scripts to keep track of the recordings, to help check if the recordings ended successfully, if they are synchronous, or to help debug an issue in case of a failure.

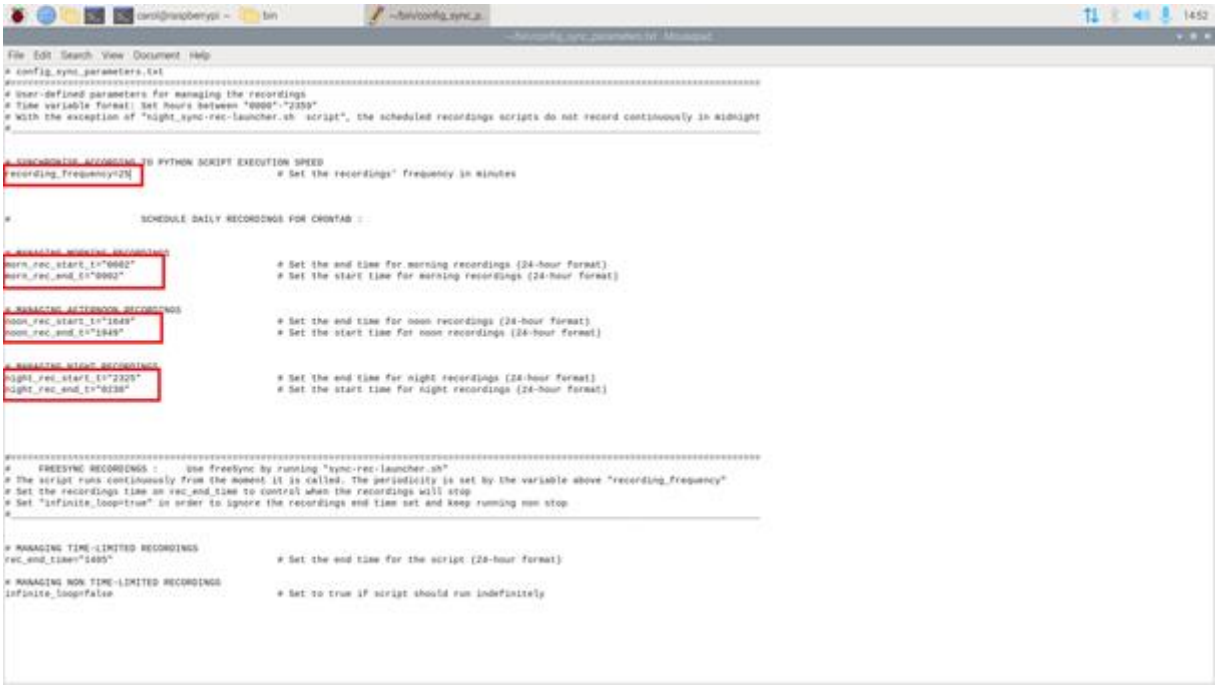


Figure 6: Recording parameters for the end user to control

The end user can control the recordings’ duration and start time, together with the periodicity, in the file “config_sync_parameters.txt” (Figure 6). These variables are essential for synchronising the recordings of the raspberries.

Summarising, when the recordings are scheduled by the end user, the system executes the following steps (Figure 7):

1. Crontab schedules bash scripts activation
2. Bash scripts run for a specific time limit, and they call the USS periodically to ensure synchronised Quadraphonic recordings
3. When called USS algorithm opens the AudioMoth mic for recordings (recording duration is set by the user)
4. USS separates the human voice from the background sounds
5. The separated human-voice is blurred in real-time
6. Background sound and blurred human voice are saved separately at the SD card, in

appropriately named directory

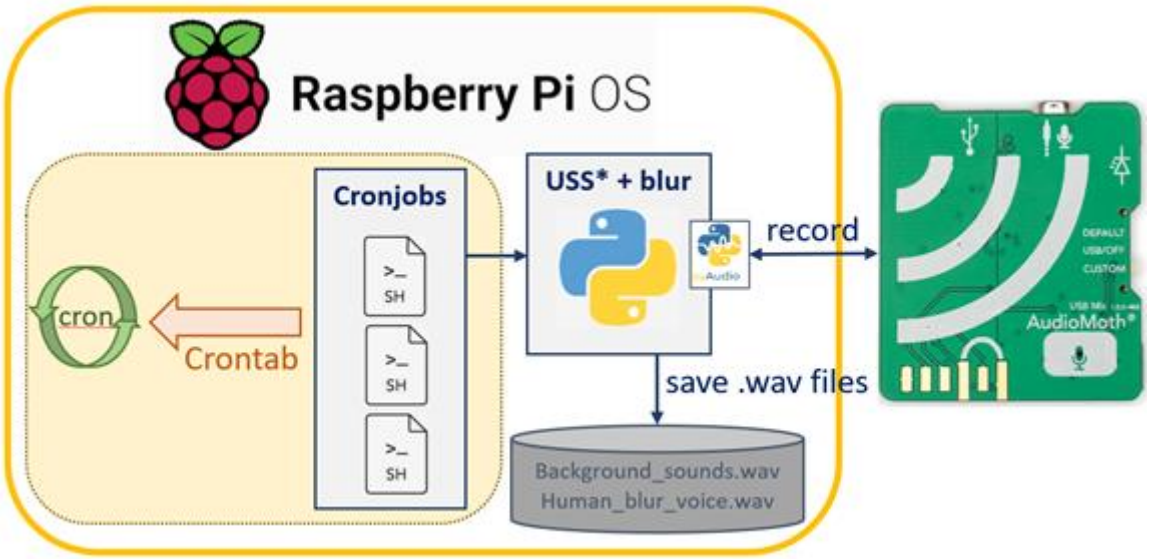


Figure 7: Software Diagram of Audio Recording Toolbox

5.2.4 **Results**

To assist the recordings of the end user it was necessary to define the combination of the values of the parameters that the user is able to set in terms of the duration for the successful execution of the scripts. Running tests have provided the timing results depicted on Table 4, that were given as guidelines to the user:

	SD card	RATE	RECORD_SECONDS	segment_seconds	Max recording duration(min)
1	64GB	44100	60	2	23
2	64GB	44100	30	2	10
3	32GB	44100	60	2	28
4	32GB	44100	60	1	27
5	32GB	44100	30	2	24
6	64GB	32000	60	2	16
7	64GB	32000	30	2	9
8	64GB	32000	20	2	7
9	64GB	32000	60	1	24
10	64GB	32000	60	1.5	18
11	32GB	32000	60	2	20
12	32GB	32000	30	2	11
13	32GB	32000	60	1	20 - 25
14	64GB	16000	60	2	10
15	64GB	16000	30	2	7
16	64GB	16000	20	2	4
17	64GB	16000	60	1	12
18	64GB	16000	60	1.5	10

19	32GB	16000	60	2	13
20	32GB	16000	30	2	10
21	32GB	16000	10	2	3

Table 4: Max recording duration according to the algorithmic variables

Performance of the Audio Recording Toolbox

Two speed classes of microSD cards were provided by the user (after our request for a faster card). A 64GB microSD card with read/write speed class 90/80 MB/sec and four 32GB microSD cards with UHS speed class 30MB/sec. The speed difference is reflected on the results of Table 4. The increase of the processing time in relation to the recording time is also significant.

Additionally, we needed to advise the artist to not use sample rate above 44100 since higher frequencies require more processing power for the USS algorithm, and the execution time was increasing significantly, leading to a balance-loss between recording intervals and processing time. Segmentation in processing has improved the times yet it has caused a deterioration in the voice separation success possibility.

Notably, the device's performance is additionally closely related to the Operating System (OS) that runs the device. Moreover, there was notable performance degradation related to the experiments conducted and software adjustments that took place for the source separation and filtering. In the end, even though system maintenance and optimisation took place, the first microSD card used to experiment on the system environment was noticeably slower than those that run with clean install.

Synchronisation response

In general, the synchronisation of the recordings is considered highly successful, since it takes place with maximum deviation of 2 secs. There is always the exception of the first interval of recordings - the first call of the script from Cron, always failing to synchronise, appearing a delay up to 30 secs. This delay is random for each scheduling and differs for each microSD card.

On site recordings have provided highly successful synchronised results. There were some recording intervals skips (for isolated and random devices) without any apparent reason from the logs, that did not affect the devices synchronisation. This fact was attributed to strong winds by the artist, that caused the AudioMoth device to lose usb connection to the Raspberry Pi devices momentarily. If there is no communication with an input device, the algorithm exits prematurely since no recording file is available for processing.

Output

As an output, two audio files in .wav format are saved to the microSD card, each with an approximate duration of 10 seconds. If the user opts to record for 60 seconds, six folders numbered 0 to 5 will be created to store each saved file. These six folders are placed inside a folder named after the precise timestamp of the recording beginning. Additionally, this timestamp-named folder is stored within a folder named by the current date, helping to organise the collected material.

Figure 8: Saved .wav files at Raspberry Pi OS

The final experiments utilised this output among others. They are described in detail in paragraph 6.4.

5.2.5 **Scalability**

The Audio Recording Toolbox was designed by taking into consideration the costs, portability, and simplicity to its usage, so that it can be also employed and prepared by people who are not technologically adept. As already described, the software architecture provides to the end user a range of capabilities on scheduling and performing recordings (run nonstop, on specific hours, while crossing midnights, shutdown, etc.), in a focused manner – by setting the parameters in three text files. The Raspberry Pi OS is similar to any desktop OS in basic functions therefore the environment may differ to what the end user might have encountered, yet it is not alien.

Moreover, the portable size also renders the Audio Recording Toolbox easy and discrete to use. Caroline Clauss intends to take advantage of these characteristics of the devices. Future considerations on mapping the sonic space of a neighbourhood as expressed by the artist, include equipping the residents of an area with the Audio Recording Toolbox and allowing them to wander in the area to capture the timbre of the region's open space.

Although the Audio Recording Toolbox has been designed for ambisonic recordings taking account synchronisation of four devices, that does not prevent the end user to use as many devices desirable to perform equivalent recordings (mono, stereo, etc.) according to the running needs or imagination.

The device comes with two tutorials that refer to people with basic technological background. The first tutorial provides easy step by step guidance to prepare the software environment by installing the necessary applications and packages and preparing the system set up, for everyone to acquire the Audio Recording Toolbox. The second Tutorial provides step by step guidance on how to schedule recordings.

5.3 **Conceptual Frameworks for Interactive Sonification to support Andrea Cera's Moving Soundscapes**

In the following subsection we describe the conceptual framework developed by UNIGE, and the related software modules developed and adopted for the artistic project of Andrea Cera.

5.3.1 **Artistic installation: relation with the scientific experiments**

Andrea Cera designed and presented preliminary results in public presentations, installations, and contributed to a pilot study started in 2024, in particular to an experiment on the impact of different types of sonification in the performance of joint emotional tasks in a dyad. This experiment produced a dataset of motion capture data of a number of dyads performing the

task under different conditions, recorded at the Casa Paganini/UNIGE premise, ad was carried out using the newly created dataset of soundscapes (intrusive, neutral, non-intrusive) that were present while participants were performing the experiment (without any explicit information about the presence of a soundscape). Currently we are analysing the data on measures of joint activities of the participants, to evaluate the impact of sonification on the quality of the task. In particular, the hypothesis is that an annoying, intrusive soundscape will cause a decrease of individual as well as joint movement quality.

5.3.1.1 Moving Soundscapes

Starting from a curiosity about sonic intrusiveness in degraded urban soundscapes, the installation "Moving Soundscapes" is a small-scale model, an abstract representation of a state of intrusiveness. A description of the artistic and research concepts follows:

_SOUND INTRUSIVENESS AS CONSEQUENCE OF HUMAN ACTIVITY

Degraded urban soundscapes can be interpreted as metaphors of a human tendency to intrude in the environment to transform it. The content of a degraded urban soundscape is the sum of thousands of sounds made by tools and machines created by humans, used by other humans, in order to pursue goals, solve problems, face challenges, fight difficulties, chase dreams;

_SOUND INTRUSIVENESS AND ITS EFFECTS ON THE BODY

The scientific literature shows that our bodies are unconsciously affected by sonic intrusiveness. In the case of a city soundscape, it is as if the degraded environment reacted back to our intrusion, in the most astute and intangible way. The scientific experiments with InfoMus-CasaPaganini/UNIGE represent further corroboration of these findings, exploring how intrusive soundscapes influence the quality of our movements;

Figure 9: Moving Soundscapes experimentation – Interaction with the installation

_MOVING SOUNDSCAPES

"Moving Soundscapes" is a cubist representation of this state of intrusiveness, and of its consequences, a skewed painting where visitors unknowingly replay intrusive scripts against a sentient computer environment.

Research Elements

_INTRUSIVENESS IN "Moving Soundscapes": SOUND

High- and low-intrusiveness sounds are at the core of the installation's audio material and derive from the [36 + 36] sounds used in the scientific experiments. High-intrusiveness sounds are captured along high-traffic roads, parkings, busy crossroads, where the human ear is aggressed by a soundscape randomly shaped by engines and mechanisms. The low-intrusiveness sounds come from the techniques used to lower the intrusiveness of the high-intrusiveness ones, including the addition of recordings of trees, wind, animals, captured in hills and valleys partially separated from the city, but still receiving some urban acoustic signature. In the installation these sounds are played back through the technique of granulation, which adds the possibility to alter several timbral and temporal dimensions. See below to understand how these sounds are mapped to imagery and user movement.

_INTRUSIVENESS IN "Moving Soundscapes" IMAGERY

The visual aspect of the installation relates to the sonic aspect. The core of the visual material is a collection of [36 + 36] images generated by CERTH with a trained multi-GAN: each image was labelled with the corresponding sound input. The mapping used to generate the images connects on one hand the spectral and timbral features of the sounds, and on the other hand a series of conditions for the image synthesis (saturation, contrast, brightness, symmetry, patterns, and natural/industrial labelling). In other words, the images are a visual translation of the intrusiveness factors of the sound (more details provided in paragraph 6.4).

A small number of photos was used to train the model, including images of active vs. abandoned industrialised zones: buildings, roads, windows, skies, trees, leaves, patches of grass, puddles. Traces of industrial human intrusion on nature, and traces of nature intruding back on abandoned human artefacts, coming back to destroy ancient signs of humans. Andrea Cera took these images in the zone where he lives (the North-East of Italy), in which urban, industrial and agricultural zones are inextricably interwoven.

_INTRUSIVENESS IN "Moving Soundscapes" INTERACTION

The relation between visitors and installation is also based on intrusiveness. If nobody Is in the installation space, or if visitors are completely still for more than 30 seconds, the system goes in a peaceful state. This peaceful state consists in projecting visual material derived from the images produced with the least intrusive sound labels, and playing back the corresponding sounds (i.e., low-intrusiveness sounds).

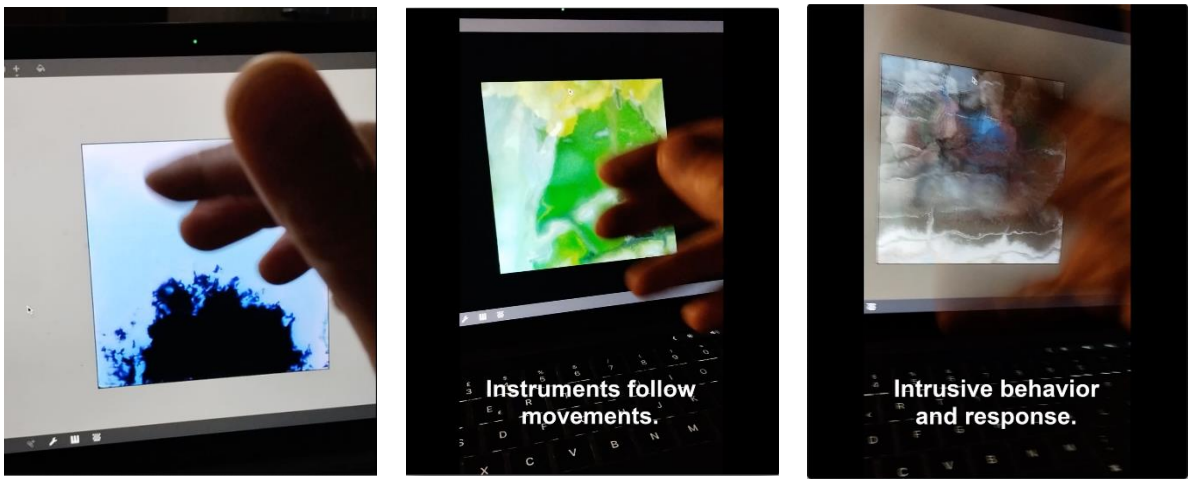


Figure 10: Moving Soundscapes hand - tests

With their movement, visitors break the homeostasis of the installation. They intrude into a form of abstract nature, and unconsciously start building abstract urban, human, artificial realities. But with the sounds created by the visitor's intrusion, the installation reacts back, intruding in the visitors' auditory system and provoking changes in movement and reactions – following the effects we studied in the scientific experiment at InfoMus-Casa Paganini/UNIGE. This means that the more disrupted, discontinuous, energetic, intrusive, violent the quality of movement is, the more the installation sound is based on high-intrusiveness audio material. The visual material changes, too: the installation starts projecting images produced with the most intrusive sound labels (the images are more contrasted, brighter, have a higher degree of industrial content, even if abstract).

This creates a feedback system, limited by the finite range of materials used by the installation, and by the amount of mental and physical energy in the visitors.

5.3.1.2 ReSilent Android app

The smartphone app ReSilent (for Android OS) is a dissemination / proof-of-concept tool created to demonstrate the low-intrusiveness sound design techniques developed in the ReSilence project by Andrea Cera. The app is intended to be activated when the user finds intrusive a certain soundscape (e.g., people talking loudly on a train trip; intrusive sounds in a hospital room; traffic outside a quiet room). Once the app has been launched, it starts transforming the sounds captured by the smartphone's microphone. The transformations are a subset of the strategies used to prepare the “low intrusiveness” version of the experiment sounds, following guidelines including timbral and temporal modification to pre-existing sounds. The processed sounds are sent to the smartphone's audio output: when listened to with headphones, these sounds blend with the intrusive sounds from outside and create a smooth mixture of real and augmented background, longer temporal shapes, a sort of abstract and fluid granulated material, which makes more difficult for intrusive sounds from outside to become salient. The programming of the internal FX chain is based on the RNBO library of MaxMSP, while the Android implementation of the RNBO patch is made using the JUCE environment. Then these steps occur:

1. The incoming audio stream is filtered and split in 3 streams: 0 to 150 Hz / 1000 to 3000 Hz. / 5000 to 8000 Hz. These three zones are not drastically and totally separated; their

boundaries slightly overlap. This allows for a certain de-correlation of output but assures that no incoming sounds gets totally ignored.

2. The 3 streams enter separated delay sections, with multiple delay lines with an amount of feedback strong enough to create tails of more than 2 seconds.
3. The same 3 filtered streams enter a section of separated envelope-followers with different slow-response times: the control signals so obtained determine the output level of six dedicated sample players, which play back the same type of “augmenting” sounds (if not the very same sounds in certain cases) used to lower intrusiveness in the scientific experiment's files.
4. The non-filtered incoming audio stream is analysed by a pitch-detection object. Every detected pitch is used as a reference note to generate a consonant trichord played back by a simple synthesiser made by three sinusoids with very slow attack and release time. The three sinusoids enter the same delay stage described in step 2.

Unfortunately, the pitch-detection used too much CPU in the smartphone, and caused audio glitches: for this reason, this particular technique was abandoned. The base-pitch generation used in the final app is based on a random generator, triggered by changes in volume of the input audio stream.

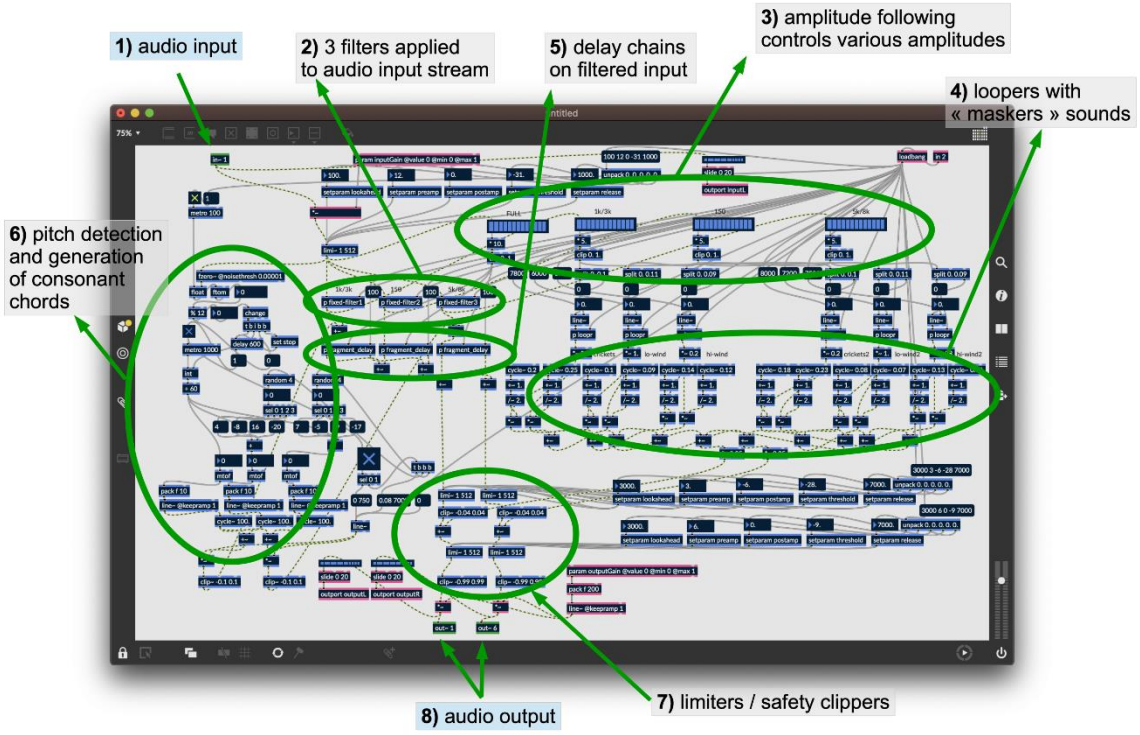


Figure 11: ReSilent app Architecture

5.3.2 **Software to support art driven experiments**

Our model of interactive sonification of human movement qualities is based on cross-modal mappings of movement qualities at different levels of abstraction and temporal scales (Camurri et al., 2005; Camurri et al., 2016). The main criteria and guidelines for our model are described in the following:

1. *Multiple Layers Cross-Modal Correspondence between movement and sound*: this is the main component of the sound architecture, and is grounded on previous research

(Camurri et al., 2005) identifying a four layers model for representing multimodal signals at different temporal scales: physical, low-, mid-, and high-levels of abstraction.

2. *Sonification integrates dynamically evolving environmental soundscapes following slowly evolving movement features related mainly to the context.* The higher-layers and temporal scales include strategies to modulate slowly varying features of background soundscapes. A sonification model at multiple temporal scales integrates simple sonic events up to complex sound streams and naturalistic simulated dynamically modulated soundscapes and modelling techniques.
3. *Slowness and continuity in dynamic curves*: movement fluidity is a mid-level quality (Camurri et al., 2016) characterised by a slowly varying response at a temporal scale in a range of half a second. The sonification of fluidity should reflect such slow pace and contribute to semi-conscious slow and fluid movements. Analogously, impulsiveness, rigidity, fragility are detected, but their mapping is subtractive, they are not sonified.
4. *Low intrusiveness sonification*: The goal of usual sound design approaches is to evoke explicit meaning, which, from an auditory scene analysis point of view, should be detectable and recognizable as clearly as possible by users. In our approach, the sonification should avoid perturbing the optimal flow during the experience, remaining at a semi-conscious level: sonification emerges without startling, intrusive, or annoying effects. This constraint influences many features of the sonification model: it raises the need for a background sound stream at a very slow temporal scale, as a stable layer through which sonification events can emerge in a controlled way, keeping under control contrasts, surprises, ambiguities. This states the necessity of carefully controlling the pitched content and reinforces the importance of smooth and slow dynamics. Our approach aims to implicitly or semi-consciously evoke, encourage, nudge, and even elicit certain qualities of movement, including fluidity, or synchronisation of behaviour in a dyad or a small group of users (Sabharwal et al., 2022). The approach is grounded on the affordance enabled by the cross-modal mapping of movement qualities to sound. That is, interactive sonification of full-body movement qualities to induce the user(s) to the desired movement qualities: for example, to change the quality of a movement in a physical exercise from a fragile or rigid execution to fluid, continuous; or to induce a dyad performing a joint action to increase their level of synchronisation.
5. *Subtractive design*: Saliency is created by subtraction, instead of addition of sound. A continuous soundscape, constantly present (slowly evolving following the user movement qualities at different temporal scales) from the beginning to the end of a session, can briefly disappear under certain circumstances, to create a salient moment and elicit attention.
6. Silence: a particular case of subtractive design. Music is the only human language including symbols for silence: the rest symbols are musical notation signs indicating the absence of a sound for a given duration. In our sonification architecture both foreground and background tend to be continuous, very soft in presence and supporting the flow of the experience but punctuated with meaningful silences. Some of these silences happen in the foreground, others happen in the background. The use of an artificial background is also aimed at improving liminality in terms of

acoustic comfort, masking unwanted noise (e.g., coming from the external).

7. Polyphony and orchestration techniques to the sonification of dyads and small groups: in case of multiple users, or of a joint movement of a human with a virtual humanoid agent, orchestration techniques as well as timbral and evolution of the harmonic content have the objective to achieve clearly perceived distinct sonification streams of separate features, as well as to design a clearly perceived unique sonification of a group feature, e.g. the sonification of the level of entrainment between two users or of a user and a virtual agent. For example, in the DanzArTe project (Camurri et al., 2024), the sonification of movements of the avatar and of the user have a clearly separated "signature".

5.3.2.1 Paul Luis' artistic project

UNIGE with the artist Andrea Cera had several online meetings with Paul Luis, on discussions preparatory to his interaction design development process for his soundscape artistic project.

5.3.3 Results

A dataset of dyads performing joint emotional tasks; future results will include a refinement of analysis algorithms for measuring qualities of individual as well as dyads movement qualities. These results will be utilised as software modules integrated in the applications developed for the public performances of Andrea Cera.

5.3.4 Future considerations and Scalability

We are currently working on movement data based on the Qualisys motion capture platform available at the Casa Paganini research centre of UNIGE. The current work includes the scalability of the algorithms to low-cost sensing technology that can be utilised in-the-wild, in particular in the setups defined in the project of Andrea Cera.

Concerning scalability and availability, we are working in the following directions:

1. Porting of EyesWeb modules as separate software modules in Python, for extending the flexibility and applicability to external applications;
2. Porting from Qualisys to low-cost video-cameras using pose-estimation software, and
3. Implementing versions on IMUs for specific movement analysis techniques.

We plan also to integrate the prototypes of interactive sonification modules developed in the Andrea Cera project.

6 AI-BASED SOUNDSCAPES

The development of the AI-based soundscapes component is the main activity of Task 3.3, which focuses on cross-modal learning to develop models capable of synthesising images from audio input and vice versa. As mentioned in D3.1, the primary goal was to generate visual representations from captured sound signals and, conversely, generate audio from images. Yet, the artistic thought, and comprehension, directs to novel approaches resulting in exploitation of more AI models and algorithms like emotion recognition, sound separation, and more. The work in the context of this task includes reviewing and expanding existing datasets that contain both sound and image data that can be used for the training and evaluation of the models employed. The derived tools developed by CERTH are aligned with ReSilence's artistic-driven approach and are tailored to the needs of each collaborating artist.

6.1 Related work to support the art driven experiments of 2nd call artists

In this session will be described the approach of the Call 2 artists that are currently exploring/investigating how to integrate AI technologies into their project. Some of these artistic projects may currently be in an initial stage of development.

6.1.1 Radioscape analysis and Story Generation for Guillem Serrahima's UBIQUITOUS NOISE project

As mentioned on paragraph 5.1.1, in the framework of ReSilence Guillem Serrahima's project focuses on the radio frequency spectrum. Currently, the artist examines two approaches. The first one is already described above, and centres around mapping radio waves, a concept inspired by radio astronomy.

The second, is story generation (pasting images in consequence) controlled by electromagnetic signals that will be modified / configured to frequencies that can be detected by the human ear. On one hand, story generation will use the electromagnetic signal captured in the urban space using the equipment presented in 5.1.1 which comes from various sources, in an effort to create a video that will consist of images correlated to media. On the other hand, this approach concerns a story generation that uses the deep space radio signals to paste images coming from optical telescopes.

At this stage the artist cannot provide specific description about the datasets that will be used thus, his artistic vision is a bit abstract. There is an arrangement with the astronomical observatory of Green Bank (West Virginia, America) which the artist will visit to record and shoot. There are discussions for acquiring datasets from various sources:

1. Radio Astronomical images provided by the astronomical observatory
2. Radio Signals provided by the astronomical observatory
3. Image datasets obtain from various media of communication (television/radio)
4. Radio signal recording, using a prototype device designed in the framework of ReSilence

The methodology that is going to be implemented mainly depends on the datasets that will be collected, and on the artist's decisions about the final project.

collaborating artist.

6.1.1.1 Story Generation

There have been various works intending to perform generation of visuals from audio or the opposite[13] (Arcand et al., 2021), or even classification, yet there are not many projects that focus on story generation. Most existing story generation approaches that apply AI techniques use text to produce the story. In a recent work Ganguly et al. (2023) used a lyrics-to-image approach, in which a set of images are retrieved corresponding to each line of the song which are automatically inserted and aligned into a video track.

When it comes to music and image correlation there are two main approaches. On the one hand there is a project that aims to connect Music and Image. Wu et al. (2016) proposed the Cross-Modal Ranking Analysis (CMRA), which uses non-linear models to learn the relationship between music and images. Their work also employs techniques like kernel-based ranking frameworks and integrates semantic representations for both music and images using features like Mel-Frequency Cepstral Coefficients (MFCC) and lyric-based attributes for image representation. These methods help optimise the matching of music and image pairs while maximising the similarity within a unified space. The benchmark dataset produced by this work used to be open source, yet the provided link is unfortunately broken.

The second approach that is encountered more often is to try to pair images to sound via emotion. Zao et al. (2020) propose the Cross-Modal Deep Continuous Metric Learning (CDCML), which is designed to match images and music based on their emotional content within the Valence-Arousal (VA) space. CDCML operates by learning a shared latent embedding space that captures emotional similarities between the two modalities, ensuring that both image and music data are aligned in this space according to their emotional attributes. By doing so, it enables cross-modal similarity prediction while also maintaining accurate emotion predictions within the individual domains.

The authors connect this method to Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models. CNNs are used primarily for emotion recognition tasks in images by extracting low-level features, such as colour and texture, and connecting them to emotional responses. In the context of music, RNNs, and especially LSTMs, are leveraged to capture the temporal dynamics of music, which is crucial for modelling how emotions evolve over time in a musical piece. The combination of CNNs and RNNs allows the model to automatically extract meaningful features from both images and music, which are then processed within the CDCML framework to predict emotion-based similarities and maintain consistency across modalities

In a different approach, a model was trained using a deep neural network architecture designed to learn affective correspondence between music and images (Verma et al., 2019). It consists of two subnetworks: one for processing visual data and another for audio data, both of which project their respective modalities into a shared representation space. This common space allows for the comparison of emotional content across modalities. During training, the network performs a binary classification task to predict whether a given music-image pair has true emotional correspondence (i.e., evokes similar emotions) or false correspondence (i.e., evokes different emotions). The model is trained on a large-scale database annotated with emotion labels—positive, neutral, and negative—to facilitate learning and evaluation. The

[13] https://chandra.si.edu/sound/#learnmore

training process involves optimising the network to accurately classify the emotional alignment of music-image pairs, and the learned representations are also utilised for emotion recognition within each modality independently.

6.1.1.2 Mapping electromagnetic waves

Regarding mapping Electromagnetic frequencies, when it comes to mapping the night sky, for decades astronomers analyse the signal using open-source software packages[14,15,16] that is traditionally used to turn radio observations into images. This software can accept as an input the signal collected from a radio telescope, which is a group of dishes placed in various places on Earth that all look simultaneously at the same point in the sky. Radio telescopes usually have complex structure including array setups with multiple antennas and/or focal plane arrays. For single dish, no array included, amateur prototypes like what is probably going to be developed in the framework of ReSilence, existing work provides more relevant solutions.

More relevant works to the data Guillem Serrahima has described as an input, is SkyNet. SkyNet is an algorithm for single-dish radio mapping (Martin et al., 2019). Instead of traditional weighted averaging, which can blur data, it uses weighted modelling to interpolate signal measurements, preserving instrumental resolution. It applies local modelling to differentiate astronomical signals from instrumental and environmental noise, and to separate signals from radio-frequency interference (RFI), even with just continuum data. The algorithm does not require data to be collected on a rectangular grid or regridded, allowing for flexible data integration and processing from multiple observations with varied pixel densities. These techniques are being integrated into the Skynet Robotic Telescope Network and the Green Bank Observatory's radio telescope to enhance the accuracy and customisation of data products.

In a most recent implementation that focuses on mapping wireless signal coverage, SpectrumNet improves radio mapping by leveraging generative AI to create high-resolution, three-dimensional (3D) radio maps (Zhang et al., 2024). Unlike previous models that are limited to two-dimensional (2D) maps and specific frequency bands, SpectrumNet encompasses 3D spatial data across five frequency bands, and includes diverse terrain and climate scenarios. This approach addresses the limitations of 2D models by incorporating environmental factors like terrain variations and climate conditions, leading to more accurate and generalizable radio maps. The dataset's breadth and detail make it suitable for studying and mapping wireless signal coverage in varied and complex environments, not just in space.

6.1.2 Harmonic melody blending for Ari Benjamin Meyers's INVINSIBLE CHOIR project

The Invisible Choir is a geolocated, AI-moderated composition that invites public participation to reshape a soundscape based on spontaneous contributions. Participants in an indoor park and an outdoor gallery will be able to contribute musical expressions using their voices and bodies via a mobile app. These public inputs—whether harmonious or dissonant—will blend with a pre-existing musical soundscape composed by the artist and the collaborating SME, forming a dynamic and ever-evolving composition.

AI will play a central role in moderating and harmonising these contributions. It will preserve

[14] https://casadocs.readthedocs.io/en/stable/
[15] https://kernsuite.info/
[16] List of software packages: https://www.atnf.csiro.au/computing/software/index.html

the uniqueness of each contribution while ensuring the overall composition remains listenable and engaging. By experimenting with AI models, the project will explore how machine learning can act as a benevolent mediator of human musical expression.

6.1.2.1 Movement Sonification - Geolocation

In addition to vocal inputs, the Invisible Choir aims to experiment with movement sonification as a tool for creating an interactive performance narrative, where participants' physical locations dynamically shape the evolving soundscape. Although the specific implementation is still under development, one possibility is the use of platforms like Roundware[17], an open-source system for geo-located audio collection and playback. If implemented, Roundware would allow participants to record and contribute audio through their mobile devices, with their geographic coordinates serving as key parameters that influence the composition. As participants move through the indoor park and outdoor gallery, their movements could alter the soundscape in real time, making location a driving factor in the unfolding performance. This approach would turn physical interaction into a creative element, narrating the soundscape through spatial presence.

The geolocated nature of the project adds another layer of complexity: the soundscape will change based on the location of participants within the park and gallery, with coordinates serving as a parameter in the AI model's training. This project ultimately asks whether AI can foster harmony in diversity, allowing for both dissonance and unity within a shared musical space.

6.1.2.2 AI Integration with MusicGen and AudioLM

For the Invisible Choir project, a combination of advanced AI music generation models, MusicGen (Copet et al, 2024) and AudioLM (Borsos et al, 2023), is proposed, to integrate precomposed music with public contributions. MusicGen is a controllable music generation model that uses both musical and textual inputs to generate high-quality compositions. In this project, MusicGen will blend the pre-composed music by the artist and his SME collaborators with real-time, short musical phrases contributed by participants. Each participant's input will be converted into textual form, allowing MusicGen to integrate it harmoniously (or dissonantly) with the precomposed base, ensuring the evolving soundscape stays coherent.

On the other hand, AudioLM excels at converting audio into text-like representations and generating structured, high-quality sound. In Invisible Choir, AudioLM will transform participant contributions into textual inputs that can be used by MusicGen. This ensures that each short musical phrase can be processed and incorporated seamlessly into the ongoing composition, maintaining both the structure and flow of the evolving soundscape.

In conclusion, the combined use of these two models is going to enable the integration of precomposed music with spontaneous public contributions, ensuring a dynamic yet coherent soundscape that evolves in real-time.

6.1.3 An interactive stage-performance for Alexander Hackl's & Wen Liu's UNCANNY REVERIE project

"Uncanny Reverie" is an interactive, AI-driven musical-theatrical journey that intertwines

[17] https://roundware.org/

climate narratives with audience participation to create a dynamically evolving performance. This project harnesses state-of-the-art AI technologies to weave climate data into both music generation and visual synthesis, creating an immersive, real-time experience.

In the study of music generation, models like AudioLM (Borsos et al, 2023), MusicGen (Copet et al, 2024), and JEN-1 (Li et al., 2024) will be explored for their distinct capabilities. AudioLM excels in producing long-term, emotionally coherent audio that captures both the structural and emotional aspects of climate data. MusicGen prioritises high-fidelity stereo output, offering precise control over musical elements through textual or melodic prompts. JEN-1, with its diffusion-based approach, adds flexibility by enabling tasks like music inpainting and melodic extension. Climate data, such as temperature anomalies and sea level rise, will be normalised and mapped to musical parameters like tempo, dynamics, and harmonic progressions. The models will be trained to interpret this mapping, enabling the generation of compositions that reflect environmental changes while allowing for controlled experimentation in the music produced.

Similarly, in the study of image generation, models such as StyleGAN, Latent Diffusion Models (LDMs) (Rombach et al., 2024), and a multi-conditional Generative Adversarial Network (GAN) will be explored, each tailored to produce visuals based on climate data features like temperature anomalies and sea level rise. LDMs are advanced generative models designed to create high-quality images efficiently by operating in a compressed latent space rather than directly in the high-dimensional image space. This approach involves first encoding images into a lower-dimensional latent representation, where diffusion processes are applied to refine the image. By incorporating conditional mechanisms, LDMs can generate images that adhere to specific attributes or constraints provided by the input conditions. The multi-conditional GAN, based on the StyleGAN architecture, will be employed to produce real-time visuals, with fine control over artistic elements such as mood, texture, and colour. These visuals will respond dynamically to climate data, ensuring that generated images are reflective of environmental changes.

The mapping of climate data to visual characteristics is still under consideration, but as already mentioned, will involve parameters like temperature anomalies and sea level rise being tied to visual elements such as saturation, brightness, and contrast. These mappings will evolve to complement the project's aesthetic goals, ensuring that the visuals reflect the environmental data as closely as the music does. This approach allows for continuous experimentation, where climate data drives both music and visuals to create a cohesive, immersive experience.

6.2 Datasets that can support the art driven experiments of 2nd call artists

This section describes the datasets most likely to be used in supporting the art-driven experiments of Call 2 artists. A detailed explanation of the datasets suited to each artistic project is provided.

6.2.1 Datasets for Guillem Serrahima's approach

There is a plethora of datasets for sound classification yet there are not many datasets that pairs image to sound. The following datasets are valid options that could be utilised in training models:

EMID: Emotionally paired Music and Image Dataset (EMID) (Zou et al., 2023) is a dataset that

consists of over 30,000 music-image pairs, each annotated with a 13-dimensional emotional model. It captures a broad spectrum of emotional states, emphasising emotional consistency between the paired music and images. Unlike traditional datasets that focus on semantic or broad emotional correlations, EMID aims to closely align the emotional content of both modalities with human perceptual understanding. The dataset's emotional labels are derived from a detailed model. Its effectiveness has been validated through a psychological experiment, which demonstrated that considering emotional relationships significantly improves matching accuracy.

IMEMNet[18]: A large-scale dataset created for emotion-based matching between images and music using continuous emotions in the VA space. This dataset is specifically designed to facilitate the study of cross-modal relationships between image and music by leveraging emotional cues, and it plays a crucial role in the experiments validating the proposed CDCML method.

IMAC[19]: Image-Music Affective Correspondence (IMAC) is a large-scale dataset designed for crossmodal emotion analysis. It consists of over 85,000 images and 3,812 songs, totalling approximately 270 hours of audio. Each sample in the database is labelled with one of three emotions: positive, neutral, or negative. The database combines these image and music samples, labelling them as having true affective correspondence if both the image and music in a pair share the same broad emotion category.

6.2.2 Datasets for Ari Benjamin Meyers' approach

The foundation of the soundscape, a long pre-composed piece of music, will provide a distinct musical structure for the AI to build upon, and it will be created and provided by Ari and the collaborating SME.

AudioSet[20]: Provides a vast collection of labelled audio events, helping the AI model recognise different types of human-made sounds and musical expressions. It consists of over 2M human-annotated 10-second video clips. These clips are collected from YouTube, therefore many of which are in poor-quality and contain multiple sound-sources. A hierarchical ontology of 632 event classes is employed to annotate these data, which means that the same sound could be annotated as different labels.

Acappella[21]: This dataset comprises around 46 hours of a cappella solo singing videos sourced from YouTube. The dataset comprises 1493 different video samples in total spanning four language categories: English, Spanish, Hindi and Others.

HumTrans[22]: A dataset that can serve as a foundation for downstream tasks such as humming melody-based music generation. It consists of 500 musical compositions of different genres and languages, with each composition divided into multiple segments. In total, the dataset comprises 1000 music segments. To collect this humming dataset, 10 college students were employed and each of them hummed every segment twice. The humming recordings were sampled at a frequency of 44,100 Hz.

[18] https://drive.google.com/file/d/1pP0iW7AGZtwzpunSTBt0A3UiNzA5tIgx/view
[19] https://gaurav22verma.github.io/IMAC_Dataset.html
[20] https://research.google.com/audioset/index.html
[21] https://ipcv.github.io/Acappella/acappella/
[22] https://cs.paperswithcode.com/paper/humtrans-a-novel-open-source-dataset-for

ComMU[23]: Dataset for Combinational Music Generation. The dataset contains 11,144 MIDI samples written and created by professional composers. They consist of short note sequences (4,8,16 bar) and are organised into 12 different metadata. Their content has the following structure: BPM, Genre, Key, Track-instrument, Track-role, Time signature, Pitch range, Number of Measures, Chord progression, Min Velocity, Max Velocity, Rhythm.

6.2.3 Datasets for Alexander Hackl's & Wen Liu's approach

MusicCaps[24]: The MusicCaps dataset contains 5,521 music examples, each of which is labelled with an English aspect list and a free text caption written by musicians. The text is solely focused on describing how the music sounds, not the metadata like the artist-name. The labelled examples are 10s music clips from the AudioSet dataset (2,858 from the eval and 2,663 from the train split).

GHCN-Daily/GHCNd[25]: The Global Historical Climatology Network - Daily dataset integrates daily climate observations from approximately 30 different data sources. Version 3 contains station-based measurements from well over 90,000 land-based stations worldwide, about two thirds of which are for precipitation measurement only. Other meteorological elements include, but are not limited to, daily maximum and minimum temperature, temperature at the time of observation, snowfall and snow depth.

GLORYS12V1[26]: This product is the CMEMS global ocean eddy-resolving (1/12° horizontal resolution, 50 vertical levels) reanalysis covering the altimetry (1993 onward). This product includes daily and monthly mean files for temperature, salinity, currents, sea level, mixed layer depth and ice parameters from the top to the bottom. The global ocean output files are displayed on a standard regular grid at 1/12° (approximately 8 km) and on 50 standard levels.

CyberVerse[27]: A captivating dataset that immerses you in the futuristic realm of cyberpunk. With a collection of 4,437 high-quality images, this dataset offers a comprehensive exploration of the cyberpunk genre, encompassing a wide array of captivating elements. From breathtaking cityscapes and intricate architectural wonders to cutting-edge gadgets, intriguing characters, and awe-inspiring robots, "CyberVerse" showcases the essence of a dystopian future intertwined with advanced technology.

6.3 AI-Based soundscapes for Caroline Clauss' SONIC DRIFT project

Caroline Clauss's project is an ADE, that explores the relationship between changing sonic environments and the transformation of urban spaces. This approach aspires to rethink the acoustic horizon through flatline constructs; thus, it focuses on disruptions in sonic patterns as they interact with dynamic city landscapes.

"Flatline constructs" refer to the homogenisation and loss of diversity in sonic environments. Murray Schafer (1977) describes them because of noise pollution and mechanical sounds that suppress natural soundscapes, leading to a monotonous auditory experience. Mark Fisher (1999) extends this idea to societal decay, where organic and inorganic elements converge,

[23] https://pozalabs.github.io/ComMU/
[24] https://www.kaggle.com/datasets/googleai/musiccaps
[25] https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00861/html
[26] https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_001_030/description
[27] https://www.kaggle.com/datasets/cyanex1702/cyberversecyberpunk-imagesdataset

creating a static, lifeless atmosphere. He calls this the "gothic flatline", reflecting a haunting sense of stillness and loss of future potential. Both concepts highlight the erosion of vibrancy in sound and experience, contrasting with Shaviro's (2010) idea of “alluring figurations”, which represent dynamic, affective points of intensity in ever-changing environments.

Using an AI-driven sonic-geographic dérive, the project maps these changes, particularly in the northern area of Chaussée d'Anvers (CDA) in Brussels, with the goal of highlighting how sound-based approaches can inform urban redevelopment strategies. The concept “AI-driven sonic-geographic dérive” - introduced in Caroline Clauss’s Sonic Drift - is a method that combines the Situationist practice of *dérive*—drifting through urban spaces to experience and analyse their dynamics—with AI technology to explore the shifting soundscapes of transforming urban environments. In this context, the dérive focuses on “*sonic spatial shifts”*, capturing how various sound patterns, or "*sonic figurations*", reveal disruptions or allure within the urban landscape.

The result of this analysis is a sonic map that captures shifts in sonic spaces and their emotional impacts in a rapidly changing urban environment. Three transformative urban spaces were selected to be analysed:

- the President's Garden, a privatised public green area at the intersection of CDA and Helihaven Avenue
- an industrial site in the northern Helihaven area currently undergoing regeneration
- and the junction square where CDA meets Masui Street.

6.3.1 Algorithms supporting the art driven experiments

To support the Sonic Drift project, models and algorithms based on machine learning were used and developed, as well as algorithms for implementing appropriate filters to retain frequencies of interest for analysis. More specifically, the USS algorithm for the separation of sound sources (also used in Audio Recording Toolbox - q.v. 5.2.3) was utilised to separate human voices from other sources during the recording of the experiments and, during the analysis of the recordings, to identify the sound sources present in them so that the sound components comprising the soundscape could be captured.

A soundscape emotion recognition model was also developed, based on the Audio Spectrogram Transformer model. The model was pre-trained on the AudioSet dataset (Gemmeke et al., 2017) and fine-tuned on the Emo-Soundscapes dataset (Fan et al., 2017) to produce Valence and Arousal values. Thus, the model takes a soundscape recording as input and outputs the perceived values for Valence and Arousal, allowing for an assessment of the comfort of the space.

For emotion recognition, several models were trained, including AST (Gong et al., 2021), BEATs (Chen et al., 2022), CNN14_16k (Kong et al., 2019), and SVM. These models were initially trained from scratch on the Emo-Soundscapes dataset for the regression tasks of Valence and Arousal, incorporating appropriate output layers. They were then fine-tuned on Emo-Soundscapes after being trained on the AudioSet dataset for classification tasks, to assess their ability to leverage knowledge from this extensive dataset. The models were trained with a 90% training and 10% test split, using the Adam optimizer. The learning rate was adjusted within the range of [0.01 - 0.0001]. Among these models, the AST model demonstrated the best performance when fine-tuned on the Emo-Soundscapes dataset following its initial training

on the AudioSet dataset.

Model	Arousal	Valence
	R^2	
AST	0.9243	0.7289

Table 5: Evaluation of the performance of the AST model

Additionally, given that all recordings included background noise from the urban environment, an algorithm based on Gaussian Mixture Models (GMM) was developed. The algorithm segments each recording into intervals based on shared statistical characteristics, such as mean, standard deviation, skewness, and kurtosis. By comparing the signal power in each interval to a predefined threshold, the algorithm determines whether the interval contains background noise or meaningful information.

Furthermore**,** two filters were implemented, a low-pass and a high-pass filter, to investigate the high-frequency noise as well as the low-frequency noise present in the soundscapes.

Finally, an appropriate algorithm was implemented to utilise the results of the models mentioned above to perform the necessary analysis and produce outcomes based on the metrics used to evaluate the soundscapes.

6.3.1.1 **Datasets**

The models used on the Sonic Drift Project to interpret the allure of a sonic figuration were the USS and the trained Audio Emotion Recognition (AER) model based on the AST, as mentioned before. These models were trained on two datasets, namely “AudioSet” and “Emo-Soundscapes” respectively.

In more detail, the separation of audio sources constitutes a fundamental problem in the scientific community, aiming to distinguish the various sources within a monophonic audio file. For this purpose, we used a powerful model (the USS) trained on a large dataset called AudioSet, with the ability to differentiate hundreds of sound classes, specifically 527. This model has the capability to separate sources into a hierarchical structure, which the AudioSet dataset follows. AudioSet dataset comprises 2 million human-annotated videos, each lasting 10 seconds. It is organised, as mentioned previously, in a hierarchical structure, covering a range of everyday sounds and sounds present in urban environments.

On the other hand, the AER model was trained on the EmoSoundscapes dataset. EmoSoundscapes consists of 6 categories of sounds: natural sounds, human sounds, sounds and society, mechanical sounds, quiet and silence, sounds as indicators, with 100, 6-second-long samples for each category, and a total of 600 samples. Additionally, it includes 613 samples that are a mixture of the previous ones (thus, the total number of recordings is 1213). Generally, this dataset contains evaluations of perceived emotion for one thousand two hundred and thirteen soundscapes. These evaluations consist of two values, the "Valence" value and the "Arousal" value, creating a 2D space, as shown in the diagram below:

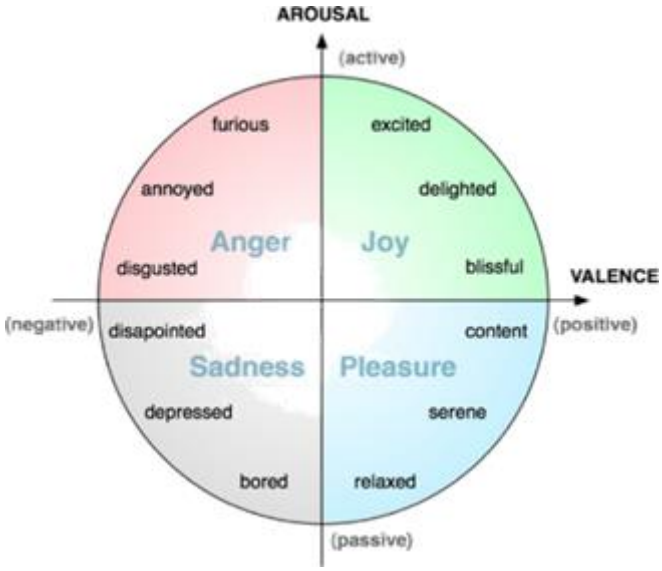


Figure 12: 2D Arousal - Valence diagram proposed by Russel

6.3.2 **AI Analysis Results**

In Sonic Drift, the AI-powered component of the dérive plays a key role in mapping these sound shifts by using models like AER and USS. These tools analyse the emotional impact of sounds—whether they induce stress or calm—and classify them by utilising soundscape ecology categories like biophony (natural sounds) or technophony (mechanical sounds). By identifying "alluring sonic figurations" (sound patterns that evoke strong emotional responses), the AI system helps trace how different sounds correlate with urban redevelopment strategies.

The dérive, in this sense, serves as a way to explore not just physical spaces but also their emotional resonance. The AI-driven analysis creates a *sonic cartography*—a map that captures sound patterns visually and aurally—while evaluating how urban transformations affect emotional well-being. Through this method, AI is used to measure and represent how different sounds, from very low frequencies (subsonic) to ultra-high frequencies (ultrasonic), impact the experience of space and reflect broader changes in the city.

In the framework of ReSilence the Sonic Drift uses two main methodologies for audio data collection: a) rooftop recordings (or passive recordings), and b) street-level audio walks around the urban block under study (or active recordings). In all cases, recordings were scheduled at key transition times (sunrise, sunset, and midnight) to capture sonic shifts in relation to three urban strategies considered representative of an urban environment by the artist. The three urban strategies are linked to the three transformative urban spaces mentioned above (Garden, Helihaven, Masui). Rooftop recordings employed four kits of Audio Recording Toolbox (for more details please refer to session 5.2). Street-level recordings used a combination of advanced microphones, including a contact microphone, a shotgun microphone, omnidirectional microphones, and a high-quality audio recorder. The data, collected in early spring 2024, focused on the sonic environment of residential areas during the weekend.

6.3.2.1 Metrics definition for sonic space interpretation

In Sonic Drift, a *"sonic space"* refers to a changing spatial environment where sounds and emotions interact ("sonic affective flows"). Different positions in the space have varying levels of sound intensity, making the experience constantly shift. A "sonic space shift" takes place when there is a major change in how sound is experienced. Although sound is always present in cities, people often fail to notice it as it moves through both spaces and persons. This concept relates to the idea of "unplace," meaning an unfamiliar space that affects how we feel, even when we are not paying attention consciously. Interventions in physical spaces can impact how we hear and experience these sounds.

To interpret the recorded sonic spaces and express the shifts and effects of possible interventions in them, focus was given in defining certain metrics. The AER algorithm was utilised, to depict the arousal and Valence fluctuation of sonic spaces (called as figurations in the framework of Sonic Drift) both for active and passive recordings. Moreover, specifically for the active recordings, strenuous but fruitful discussions striving to align the Sonic Drift project with the philosophical approach of Mark Fisher lead into adopting three approaches to describe the sonic shifts. These approaches were translated into three main metrics:

1. **Stress/calm levels:** This metric was defined by the 2D VA diagram, as a level of how stressful or calming a recording is interpreted. The recording could be the original file recorded in a specific position, or an isolated sonic space (and more precisely the figuration). The stress level prediction of each figuration was calibrated in a 0 to 1 scale - from the respective values of Arousal and Valence.
2. **Sonic space presence**: The USS algorithm provided insight in the presence of sounds according to the local figuration by utilising soundscape ecology classifications (biophony, anthropophony, technophony). Separating sound sources that define a figuration and then measuring the intensity presence, produces this metric. The figuration's intensity presence is calibrated in a scale ranging from 0 to 1,414 and depicts the average duration it may appear sonicaly.
3. **Subsonic & ultrasonic substrate:** Frequency analysis aimed to detect the presence of subsonic and ultrasonic noise in LAeq (Equivalent Continuous Sound Pressure Level), that are repeatedly indicated by existing research as a cause of discomfort (Pawlaczyk-Łuszczyńska et al., 2005; Araújo et al., 2020; Fletcher et al., 2018). The subsonic (lfn) and ultrasonic (hfn) noise is measured in LAeq (of a dB scale).

6.3.2.2 Statistical analysis

As already mentioned, the statistical analysis took place for three transformative urban spaces, the President's Garden, the Northern Helihaven Avenue area, and the Masui Street, that are linked to the three urban strategies. The rooftop where passive recordings took place, acted as the sonic horizon of all these areas since it was placed in between (Figure 14b, and 14c). Figure 13 depicts the workflow for the analysis of the soundscape in terms of the alluring figurations. The audio input, passive and active, was processed and analysed as an ensemble to extract meaning (using the metrics) for each urban area. Statistical analysis on the output of the algorithms produces the cartographic depiction.

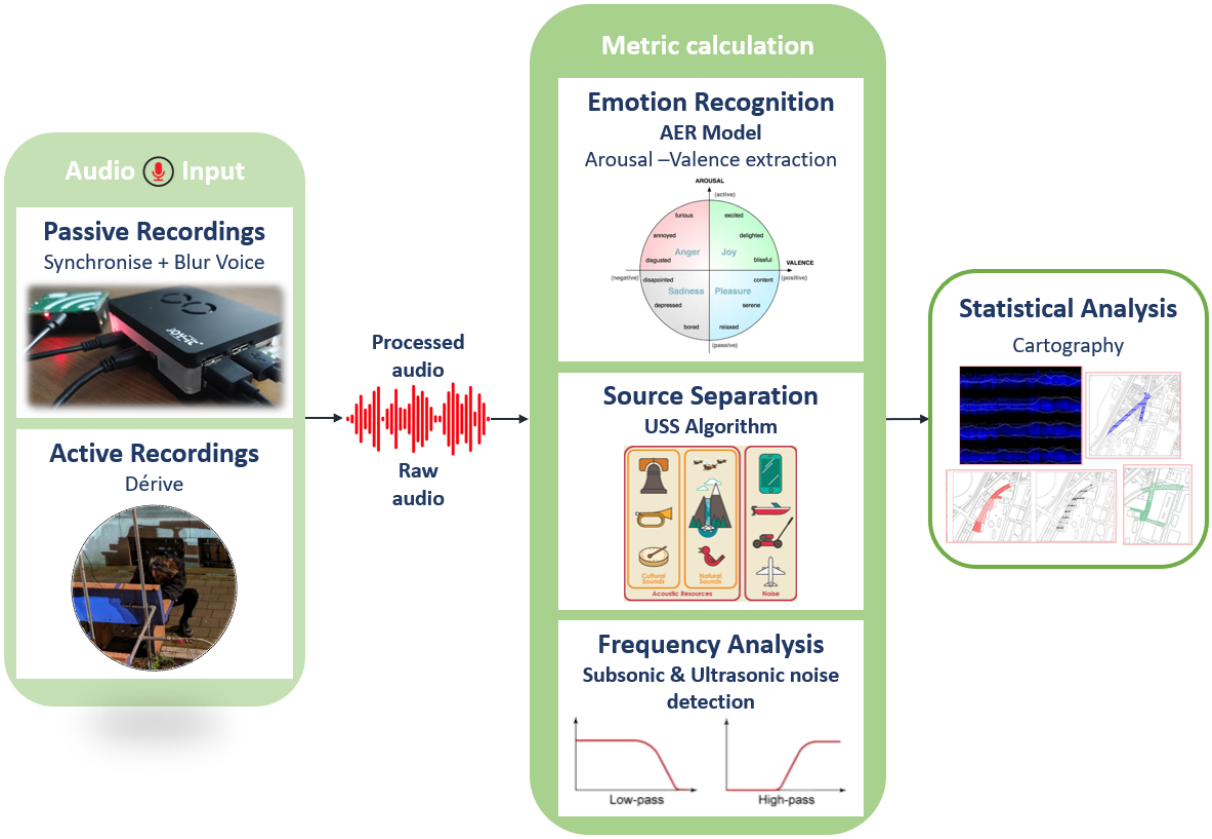


Figure 13: SONIC DRIFT workflow

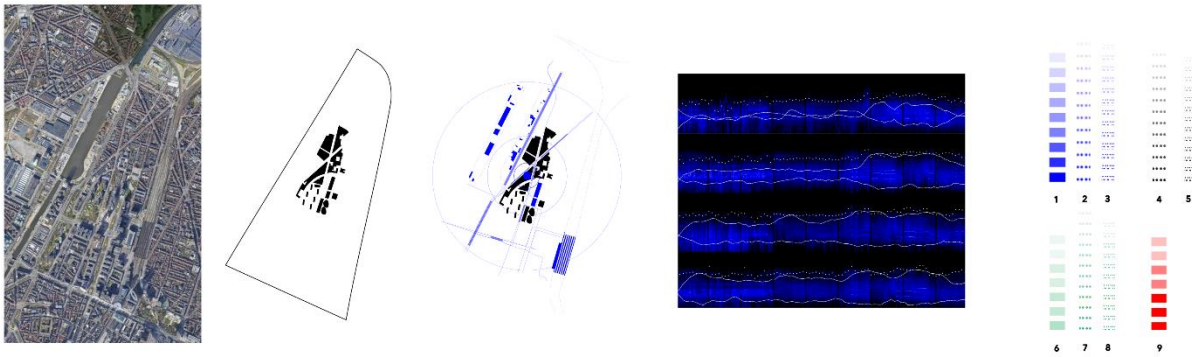
In more detail, to calculate the stress/calm levels, the AER model was trained on audio files of six seconds in duration, thus, to infer the emotional sense from longer files, a six-second window was applied to the signal under analysis, which slides by one second each time. As a result, we obtain the Valence and Arousal values for each second. Then, after identifying all the time points where the Valence and Arousal values represent the emotion of stress, the average of the Valence and Arousal values was taken. Subsequently, the average intensity of stress was calculated based on the hypotenuse formed in the VA space by the average values of Valence and Arousal.

To determine the percentage presence of the figuration in the recording, the file was first analysed using USS to identify all the sound sources that are present in the recording. Then, the sources corresponding to the specific figuration of interest were identified, and their duration within the overall signal was calculated. The ratio of the figuration's duration to the total duration of the recording results in its percentage presence.

Lastly the subsonic and ultrasonic substrate were measured as an indicator of the discomfort levels in the area. In each recording, a high-pass filter (>20 kHz) and a low-pass filter (<20 Hz) were applied, and their intensity was calculated based on the LAeq metric that represents the constant sound level that, over a given period, would produce the same energy as the actual fluctuating sound levels measured during that period.

The results of this analysis are represented in the following sonic maps, and for each area. The rooftop passive logging, as the acoustic horizon, oversees all these three areas. In an artistic depiction, spectrogram analysis of frequency, intensity was combined with the arousal and valence (upper and lower white lines respectively) fluctuation. The presence of technophony

sound sources is also depicted as dots (Figure 14d).



(a) (b) (c) (d) (e)

Figure 14 : (a) CDA North Urban Block in Brussels North District in 2024. (b) & (c) Acoustic Horizon CDA North in 2024. (d) Spectrogram analysis using passive logging of the CDA North Acoustic Horizon in 2024. (e) Legend artistic explaining

On the sonic maps the indices levels are depicted from low (faint coloured) to high (vivid coloured). Bleu indices represent technophony presence, while in dots density the low frequency noise (lfn) and high frequency noise (hfn) presence is depicted. The green indices represent the biophony presence. Red indices represent the stress levels. Finally, black indices depict in dots the density of lfn and hfn exclusive figuration of analysis.

Figures 15, 16, and 17 depict the analysis result for each urban area. A satellite picture (a) depicts the ground plan of each area. The outline of the ground plan and the chromatic portrayal represents the geographic variation of the metrics, for each recording area surrounding the alluring figuration. Initially, the presence of the sonic figuration in the public space surrounding the sonic space is given in percentage % of audible / non audible at (b). Stress Level in the presence of the figuration is portrayed at (c). Presence of non-audible frequencies in the absence of the figuration at (d). The final cartographic depiction portrays the Stress Level in the absence of the figuration (e).

(a) (b) (c) (d) (e)

Figure 15: Cartographic portrayal of the metrics for the President's Garden surrounding area

(a) (b) (c) (d) (e)

Figure 16: Cartographic portrayal of the metrics for Helihaven Avenue

(a) (b) (c) (d) (e)

Figure 17: Cartographic portrayal of the metrics for Masui Street

The following table assembles the statistical results that produced the above cartographic depictions:

		figuration absent			figuration present			
	Position	Stress Intensity	lfn substrate (dB)	hfn substrate (dB)	Sonic space presence (%)	lfn substrate (dB)	hfn substrate (dB)	Stress Intensity
President's Garden	2	0.436	55.033	5.219	0.461	-64.750	9.330	0.000
	4	0.305	47.355	-4.110	0.518	-62.535	-3.921	0.048
	5	0.066	48.933	-2.958	0.340	-61.352	-5.442	0.164
	6	0.320	52.056	1.100	0.434	-61.126	1.902	0.247
	7	0.114	50.967	-11.991	0.028	-69.438	-39.479	0.000
	9	0.176	50.637	4.380	0.472	-60.075	5.420	0.066
Helihaven	2	0.235	55.169	7.023	0.521	63.909	-1.015	0.756

Avenue	3	0.444	54.451	-0.421	0.247	57.619	3.955	0.690
	4	0.457	56.332	7.010	0.318	58.507	-2.181	0.775
	7	0.493	55.193	6.224	0.273	58.003	5.559	0.766
Masui Street	1	0.035	50.099	-4.640	0.264	55.059	-2.304	0.158
	2	0.232	50.115	-4.867	0.075	55.900	-0.597	0.259
	3	0.026	48.903	-2.617	0.202	50.962	0.708	0.092

Table 6: Statistical analysis for each recording position for the three strategies

The analysis in the prism of Sonic Drift's purpose, demonstrates that urban environments can have a profound emotional impact through both audible and inaudible sounds, shaping stress levels and emotional responses in ways previously overlooked. By leveraging AI, the analysis uncovers how sounds, whether from street activity or natural elements, contribute to emotional experiences that are often untethered from the physical characteristics of space. These insights reveal the importance of considering sonic influences in urban planning, highlighting how soundscapes play a role in creating environments that either heighten or alleviate stress. The research calls for a more thoughtful integration of sound into urban design, emphasising the need to account for both the tangible and intangible effects of the acoustic landscape.

6.4 **AI-Based soundscapes for Andrea Cera's project**

The creation of AI-based soundscapes for Andrea Cera's "Moving Soundscapes" involves an intricate fusion of auditory and visual elements, generated through the interaction of sound features and machine learning techniques. Cera's approach focuses on using high- and low-intrusiveness sounds—captured from urban, industrial, and natural environments—to influence a generative model capable of translating these sonic characteristics into visual representations. Leveraging datasets of soundscapes and a multi-GAN architecture, the process begins with extracting key spectral and timbral sound features, which are then mapped to visual elements such as saturation, contrast, brightness, symmetry, patterns and nature/industrial labelling. This sound-to-image translation allows for the creation of dynamic, evolving visuals that reflect the intrusive or peaceful nature of the soundscape. The generated images mirror not only the intensity and texture of the sounds but also the complex interaction between human activity and the environment, making AI a crucial mediator in rendering Cera's vision of sound intrusiveness through a cohesive audiovisual experience. The next sections will explore the specifics of data provided, feature extraction and mapping and the model's training process.

6.4.1 Algorithms supporting the art driven experiments

This section presents the key components that support the creative process behind the AI-generated soundscapes, where the relationship between sound and visual elements is emphasised. The following subsections detail the datasets employed, the methodologies for extracting and mapping sound and visual features, and the training process of the generative model.

6.4.1.1 Datasets

For the creation of AI-based soundscapes in Andrea Cera's project, a combination of visual and audio datasets was curated to serve as the foundation for training the generative models. The visual dataset consists of 250 images provided in .jpg format, divided into three categories: 100 images corresponding to the "industrial" or "IND" category, 100 images for the "natural" or "NAT" category, and 50 images for a "mixed" or "MIX" category, where elements of both industrial and natural environments are present. In addition to these, 200 morphing images between the "IND" and "NAT" categories were generated and provided as augmentation material in .png format. All images are in a resolution of 720x720 pixels, but for model training purposes, they were resized to 520x520 pixels.

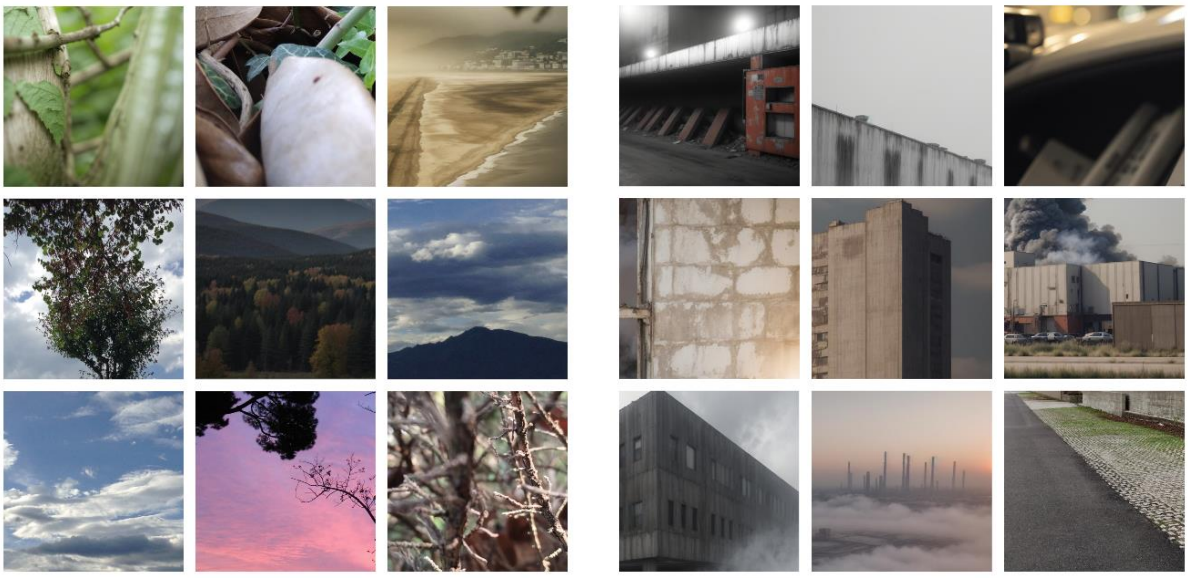


Figure 18: (Left) Images of natural labelling and (Right) of industrial labelling

The audio dataset includes a collection of 36 urban sounds, divided into two levels of intrusiveness—high and low—comprising 36 sounds in each category. Each sound file is 5 seconds long. Additionally, two 3-minute audio files were created by combining different sounds from both the intrusive and non-intrusive categories. These longer files, provided in .wav format, act as complex sonic landscapes that reflect the varying levels of sound intrusiveness. Together, these audio-visual datasets form the core material for training the multi-conditional generative model that powers the dynamic interaction in Cera's "Moving Soundscapes" installation.

6.4.1.2. Sound and visual feature extraction for Sound-to-Image mapping

Cera's project presents a system that maps specific sound features to corresponding visual characteristics, offering a synesthetic experience where audio data drives visual output. This

mapping aims to merge auditory and visual perceptions into a cohesive sensory experience. By converting sound attributes into visual qualities, the system enables participants to "see" sound, creating an interactive relationship between music and visual aesthetics.

Visual feature extraction

The process of calculating visual features from images is a sophisticated method that quantifies characteristics such as saturation, contrast, brightness, symmetry, asymmetry, and pattern density, ultimately linking these to the soundscape features in Cera's project.

Saturation, which refers to the intensity of colours in an image, is calculated by converting the image into its HSV (Hue, Saturation, Value) representation. By isolating the saturation channel, the mean value across the image is computed, providing a single metric that represents the colour intensity. This value is then normalised to a scale of 0 to 1, where 0 corresponds to grayscale or desaturated images, and 1 indicates fully saturated, vibrant colours.

Contrast and **brightness** are crucial factors in determining the tonal and luminance qualities of an image. Contrast is measured by converting the image to grayscale and calculating the standard deviation of pixel intensities, which quantifies how spread out the light and dark areas are. Higher values indicate more contrast. Brightness is similarly extracted from the HSV colour space, using the "value" channel, and represents the overall lightness of the image. Both features are scaled to a range of 0 to 1.

Symmetry is computed by comparing an image to its horizontally flipped version, assessing how closely the two match. The difference between the original and flipped images is summed to provide a symmetry score. This score is normalised based on the maximum possible value for symmetry (depending on image size) and inverted to calculate **asymmetry**.

The number of **patterns** in an image is derived from thresholding and binarization the image using Otsu's method, followed by labelling connected regions. This provides a count of distinct patterns or shapes within the image, representing the complexity or texture. These patterns are then quantized into intervals to map them onto a 0 to 1 scale, providing a simplified representation of pattern complexity.

Lastly, each image is **labelled** as either "natural" or "industrial" based on its filename, which serves as a way to categorise the visual style. Natural images are more likely to be organic and fluid in appearance, while industrial images lean towards geometric, structured patterns. This classification reflects different auditory environments, allowing for a more contextualised connection between the sound and its corresponding visuals.

Sound feature extraction and mapping

dimension	***description***	***if high, images are:***	***if low, images are:***
centroid	centre of gravity	prevalence of saturated colours	prevalence of de-saturated colours
sharpness	quality of high frequencies in the sound	high brightness/contrast image	low brightness/contrast image

skewness	distribution of frequencies around the centroid	asymmetrical distribution of objects/lines in the image	symmetrical distribution of objects/lines in the image
roughness	small fluctuations of amplitude	modulated, rhythmic, recurrent, repetitive, patterned shapes	random, non regular shapes, absence of patterns

Table 7: Suggested Sound-to-Image mapping

At the core of the mapping is the **spectral centroid**, a measure of where the "centre" of a sound's frequency spectrum lies. The centroid is mapped to saturation in the visuals, where higher frequency-dominant sounds (such as bright or high-pitched tones) lead to increased colour saturation, creating more vibrant, intense colours. This direct connection between tonal quality and colour vibrancy helps visually reflect the tonal brightness or depth of a sound. Thus, softer or lower-pitched sounds yield muted tones, while brighter sounds result in vivid, saturated visuals.

Sharpness, which relates to the amount of high-frequency content, influences brightness and contrast in the visuals. Sharp sounds with more high-frequency components make the visual output brighter and enhance the contrast between light and dark areas. As sharpness increases, the visual world becomes more defined, with sharper lines and bolder contrasts, mimicking the auditory clarity that sharpness represents. This ensures that highly defined sounds are visually echoed with equally vivid images. Sharpness is often calculated using a model based on the specific loudness distribution in the critical bands of the auditory system. The model used here is based on the Zwicker and Fastl method (Mengqiu et al., 2023).

The **skewness** of the sound spectrum, which reflects the asymmetry of frequency distribution, is tied to asymmetry in the visual depiction. When the skewness is high, the visuals become more asymmetrical, introducing irregular shapes and imbalanced compositions. Conversely, when skewness is low, the visuals tend toward symmetry, offering balanced, harmonious images.

Roughness, a measure of the perceived harshness or texture of a sound, corresponds to the complexity of patterns in the visual representation. The roughness model used is based on pairs of sinusoids with closely spaced frequencies, as modelled by Sethares (1998) and Vassilakis (2001). Sounds with higher roughness produce more intricate, textured visual patterns, suggesting a tactile element within the visuals that mirrors the texture of the sound. As roughness increases, the patterns become more complex, while smoother sounds lead to simpler, more fluid visuals. Finally, a **labelling system** distinguishes sounds as either natural or industrial, determining whether the visuals appear organic, fluid, and nature-inspired, or mechanical and geometric, reflecting urban or manufactured environments.

Once extracted, the values are normalised to allow for comparison across sounds by scaling them within a 0-1 range, where 0 corresponds to the minimum value in the dataset and 1 corresponds to the maximum. This ensures consistency in the analysis and interpretation of the results. **Skewness data has been specifically adjusted to ensure that skewness inputs can result in symmetric visualisations**, even at certain thresholds, creating flexibility in how balance is represented. The calculation is as follows: Skewness is calculated, normalised and then subtracted from 1, in order to map into symmetry.

This intricate mapping creates a rich audiovisual experience that transforms sound into visual art, allowing participants to engage with sound in a multisensory way.

6.4.1.3. **Generative model training**

Initially, StyleGAN2-ADA (Karras et al., 2020) was fine-tuned on the dataset to generate images. The results prompted the artist to clarify their vision for the desired visual outcome. However, due to the limited control over the generated image content and the need to incorporate audio features into both the content and style of the visuals, multi-conditional GANs (Dobler et al., 2022) were explored and trained. This approach provided a more structured way to align the visual output with the desired artistic and audio-driven parameters. A model that didn't result in overfitting and kept more abstraction was chosen.

6.4.2 **Results**

These are the resulted images of the trained StyleGAN2-ADA model:

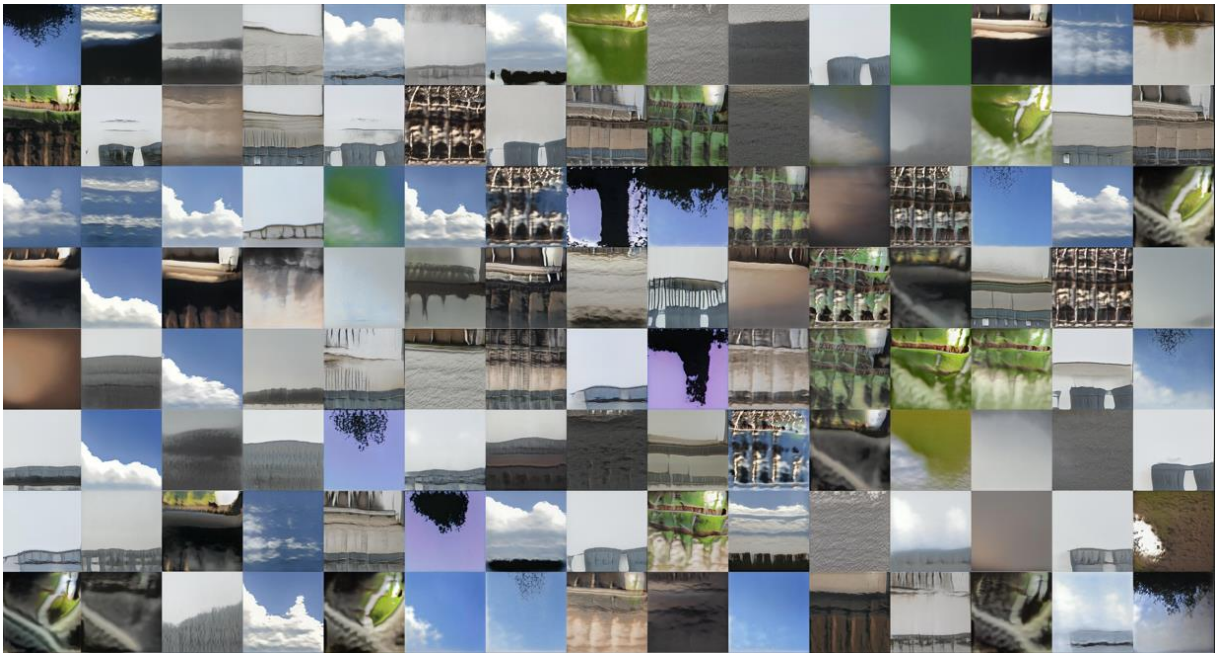


Figure 19: Generated images of the StyleGAN2-ADA

This model can generate highly realistic images based on the context it was trained on.

Some resulted images of the chosen multi-conditional GANs model are:

Figure 20: Generated images of the multi-conditional GAN

Here are images generated from the trained multi-conditional GAN, corresponding to sound input:

Figure 21: Images generated with multi-GAN from high intrusiveness (left) sound input and low intrusiveness (right) sound input

Some of the images have produced particularly interesting results (according to the feedback provided from the artist in relevance to his aesthetics and vision for the project). With the limited data available, it seems that the multi-conditional GAN model shows a slight bias toward specific conditions. However, this suggests that with more data, we could achieve even better outcomes. At this stage, the 200 images used for training seem insufficient—it is estimated that 500 to 1000 images would be more appropriate for more reliable results.

The following images demonstrate how an image can be altered by changing one condition at a time. This is an example on a picture produced with the input of the sound "4_high_intrusiveness". We also experimented with not only changing the label but also producing a mix of industrial and natural picture, as it can be shown in the last picture.

Figure 22: Generated image from sound 4 of high intrusiveness dataset with multi-conditional GAN and alterations of it by changing a corresponding visual feature value

6.4.3 **Future considerations**

In the future, developing three distinct styles of imagery is under consideration, each corresponding to different types of user movements. For instances where the system detects imbalance or dissonance, we will generate images with grainy, textured characteristics. For slow, fluid movements, the visuals will be smooth and flowing, evoking a sense of natural fluidity. Energetic or abrupt movements will be represented by angular, sharp, and defined shapes, reflecting the intensity and dynamism of the motion.

Additionally, Andrea Cera's image dataset could be expanded to ensure more precise alignment between the visual outputs and the system's movement-based inputs.

In conclusion attention should be placed in the fact that the projects of Call 2 artists that are currently in the development phase, may alter slightly - requiring additional methods or technologies to be utilised, since the dialogue between artists and researchers creates dynamic ideas that evolve and direct the projects into novel directions.

7 CONCLUSIONS

The completed and ongoing activities in the ReSilence project are demonstrating the trans-disciplinary collaboration between the scientific and technological developments provided by ReSilence partners and the artistic projects by the selected artists.

Chapter 4 accounts the development of a scalable movement analysis system, using Andrea Cera's artistic project as testbed, and involving many of the selected artists in this discussion. In alignment with ReSilence's research objectives –and more precisely with T3.1– a soundscape dataset was created containing a series of movement recordings of dyads performing joint emotional tasks, concerning both individuals or groups.

Chapter 5 contains an in-depth analysis and presentation on the equipment designed –or currently in development– to collect ambisonic soundscape recordings and movement-related soundscape data. Software, libraries, and hardware, designed to facilitate this soundscape recording process and to conduct soundscape analysis are also discussed in detail. The work is presented by artistic project for each ADE approach and meets the needs outlined in T3.2 support.

Chapter 6 focuses on the work involving AI algorithms for soundscape analysis. Tasks such as audio-to-image synthesis, emotion recognition, sound separation were employed to achieve the ReSilence;s objectives, while addressing the artistic requirements. Additionally, in accordance with T3.3 planning, multi-conditional GANs were were used to generate images from sound, incorporating multimodal features, and resulting in a dataset of 250 .jpg images and 36 urban sounds, categorised into two levels of intrusiveness.

As the ReSilence project progresses toward completion, the realisation of the projects described above will result in artistic demonstrations, more scientific and artistic experiments, showcases, and datasets. Many of these outcomes particularly those related to Call 2 artists, are still evolving and will be demonstrated in public events and scientific publications.

8 REFERENCES

Alborno, P., Volpe, G., Mancini, M., Niewiadomski, R., Piana, S., & Camurri, A. (2019). The multi-event-class synchronization (MECS) algorithm. *arXiv preprint arXiv:1903.09530*. JAM-8 Intl Workshop, Casa Paganini, Genoa

Araújo Alves, J., Neto Paiva, F., Torres Silva, L., & Remoaldo, P. (2020). Low-Frequency Noise and Its Main Effects on Human Health—A Review of the Literature between 2016 and 2019. Applied Sciences, 10(15), 5205. https://doi.org/10.3390/app10155205Burtner, M. (2003). Regenerative Feedback in the Medium of Radio: Study 1.0 (FM) for Radio Transceiver. *Leonardo Music Journal*, *13*, 39-42.

Arcand, K. K., Russo, M., & Santaguida, A. (2021, January). Chords of the Cosmos: Converting Data of our Universe into Sound. In *American Astronomical Society Meeting Abstracts* (Vol. 237, pp. 412-02).

Burtner, M. (2003). Regenerative Feedback in the Medium of Radio: Study 1.0 (FM) for Radio Transceiver. *Leonardo Music Journal*, *13*, 39-42.

Camurri, A., Volpe, G., De Poli, G., & Leman, M. (2005). Communicating expressiveness and affect in multimodal interactive systems. *Ieee Multimedia*, *12*(1), 43-53.

Camurri, A., Volpe, G., Piana, S., Mancini, M., Niewiadomski, R., Ferrari, N., & Canepa, C. (2016, July). The dancer in the eye: towards a multi-layered computational framework of qualities in movement. In *Proceedings of the 3rd International Symposium on Movement and Computing* (pp. 1-7).

Camurri, A., Seminerio, E., Morganti, W., Canepa, C., Ferrari, N., Ghisio, S., ... & Pilotto, A. (2024). Development and validation of an art-inspired multimodal interactive technology system for a multi-component intervention for older people: a pilot study. *Frontiers in Computer Science*, *5*, 1290589.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D.C., Chen, Z., & Wei, F. (2022). BEATs: Audio Pre-Training with Acoustic Tokenizers. ArXiv, abs/2212.09058.

Cohen-Hadria, A., Cartwright, M.B., McFee, B., & Bello, J.P. (2019). Voice Anonymization in Urban Sound Recordings. 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), 1-6.

Deng, J., Teng, F., Chen, Y., Chen, X., Wang, Z., & Xu, W. (2023). {V-Cloak}: Intelligibility-, Naturalness-& {Timbre-Preserving}{Real-Time} Voice Anonymization. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 5181-5198).

Dobler, K., Hübscher, F., Westphal, J., Sierra-Múnera, A., de Melo, G., & Krestel, R. (2022). Art creation with multi-conditional StyleGANs. arXiv preprint arXiv:2202.11777.

Fan, J., Thorogood, M., & Pasquier, P. (2017, October). Emo-soundscapes: A dataset for soundscape emotion recognition. In *2017 Seventh international conference on affective computing and intelligent interaction (ACII)* (pp. 196-201). IEEE.

Fierro, L., Rämö, J., & Välimäki, V. (2019). Adaptive Loudness Compensation in Music Listening.

Fisher, M. (1999). Flatline constructs: Gothic materialism and cybernetic theory-fiction [PhD thesis, University of Warwick]. http://webcat.warwick.ac.uk/record=b3252748~S15

Fletcher, M. D., Lloyd Jones, S., White, P. R., Dolder, C. N., Leighton, T. G., & Lineton, B. (2018). Effects of very high-frequency sound and ultrasound on humans. Part I: Adverse symptoms after exposure to audible veryhigh frequency sound. The Journal of the Acoustical Society of America, 144(4), 2511–2520. https://doi.org/10.1121/1.5063819

Friz, A. (2011). The Art of Unstable Radio. *Langlois, Sakolsky and van der Zon, Islands of Resistance*, 167-182.

Friz, A. (2011). *The Radio of the Future Redux: Rethinking Transmission through Experiments in Radio Art*. York University.

Ganguly, D., Parker, A., & Aji, S. (2023, March). Automatic Videography Generation from Audio Tracks. In *European Conference on Information Retrieval* (pp. 281-287). Cham: Springer Nature Switzerland.

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

Gerzon, M. A. (1973). Periphony: With-height sound reproduction. *Journal of the audio engineering society*, *21*(1), 2-10.

Glinsky, A. (2000). *Theremin: ether music and espionage*. University of Illinois Press.

Gong, Yuan & Chung, Yu-An & Glass, James. (2021). AST: Audio Spectrogram Transformer. 571-575. 10.21437/Interspeech.2021-698.

Hall, M. (2018, July). The legacy of free radio on contemporary radio arts practice. In *The Radio Conference–A transnational Forum 2018* (pp. 1-18).

Hill, A. P., Prince, P., Snaddon, J. L., Doncaster, C. P., & Rogers, A. 2019. "*AudioMoth: A lowcost acoustic device for monitoring biodiversity and the environment",* HardwareX, 6, e00073.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. Advances in neural information processing systems, 33, 12104-12114.

Kogawa, T. (2008). Radio in the chiasme. In Heidi Grundmann et al (Eds.), Re-inventing radio: Aspects of radio as art (pp. 407-409). Frankfurt Am Main: Revolver

Kolykhalova, K., Gnecco, G., Sanguineti, M., Volpe, G., and Camurri, A. (2020) Automated analysis of the origin of movement: An approach based on cooperative games on graphs. IEEE Transactions on Human-Machine Systems, 50(6), 550 – 560. DOI: 10.1109/THMS.2020.3016085

Kong, Qiuqiang & Cao, Yin & Iqbal, Turab & Wang, Yuxuan & Plumbley, Mark. (2020). PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. 10.48550/arXiv.1912.10211.

Li, P. P., Chen, B., Yao, Y., Wang, Y., Wang, A., & Wang, A. (2024, June). Jen-1: Text-guided universal music generation with omnidirectional diffusion models. In 2024 IEEE Conference on Artificial Intelligence (CAI) (pp. 762-769). IEEE.

Maheshwarappa, M. R., Bowyer, M. D., & Bridges, C. P. (2017). Improvements in CPU & FPGA

performance for small satellite SDR applications. *IEEE Transactions on Aerospace and Electronic Systems*, *53*(1), 310-322.

Martin, J. R., Reichart, D. E., Dutton, D. A., Maples, M. P., Berger, T. A., Ghigo, F. D., ... & Harvey, J. (2019). Skynet algorithm for single-dish radio mapping. I. Contaminant-Cleaning, mapping, and photometering small-scale structures. *The Astrophysical Journal Supplement Series*, *240*(1), 12.

Mengqiu, Zhu & Yu, Lingjie & Wang, Zongbiao & Ke, Zhenxia & Zhi, Chao. (2023). Review: A Survey on Objective Evaluation of Image Sharpness. Applied Sciences. 13. 2652. 10.3390/app13042652.

Niewiadomski R., Mancini M., Cera A., Piana S., Canepa C., Camurri A. (2019) Does embodied training improve the recognition of mid-level expressive movement qualities sonification?. Journal on Multimodal User Interfaces, v. 13, n. 3, pp. 191--203, ISBN/ISSN: 1783-8738, Sep, 2019.

Panda, A. R., Mishra, D., & Ratha, H. K. (2014). Fpga implementation of software defined radio-based flight termination system. *IEEE Transactions on Industrial Informatics*, *11*(1), 74-82.

Pawlaczyk-Łuszczyńska, M., Dudarewicz, A., Waszkowska, M., Szymczak, W., & Śliwińska-Kowalska, M. (2005). The impact of low frequency noise on human mental performance. International Journal of Occupational Medicine & Environmental Health, 18(2).

Rahman, M. H., & Islam, M. M. (2016). A practical approach to spectrum analyzing unit using rtl-sdr. *Rajshahi University Journal of Science and Engineering*, *44*, 151-159.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

Sabharwal, S. R., Varlet, M., Breaden, M., Volpe, G., Camurri, A., & Keller, P. E. (2022). huSync-A model and system for the measure of synchronization in small groups: A case study on musical joint action. *IEEE Access*, *10*, 92357-92372.

Schafer, R. M. (1977). The tuning of the world (1st ed). A. A. Knopf.

Sethares, W. A. (1998). Tuning, Timbre, Spectrum, Scale. London: Springer-Verlag.

Shaviro, S. (2010). Post-cinematic affect. 0 [zero] Books.

Soto-Sanfiel, M. T., Freeman, B. C., & Angulo-Brunet, A. (2022). Understanding radio art reception. *Profesional de la información*, *31*(4).

van der Heide, E. (2000) “Radioscape”, immersive electromagnetic environment, see <http://www.evdh.net/radioscape/>, accessed May 2024

Vassilakis, P. N. (2001). Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance. Doctoral Dissertation. Los Angeles: University of California, Los Angeles; Systematic Musicology.

Verma, G., Dhekane, E. G., & Guha, T. (2019, May). Learning affective correspondence between music and image. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3975-3979). IEEE.

Wu, X., Qiao, Y., Wang, X., & Tang, X. (2016). Bridging music and image via cross-modal ranking

analysis. *IEEE Transactions on Multimedia*, *18*(7), 1305-1318.

Zhang, S., Jiang, S., Lin, W., Fang, Z., Liu, K., Zhang, H., & Chen, K. (2024). Generative AI on SpectrumNet: An Open Benchmark of Multiband 3D Radio Maps. *arXiv preprint arXiv:2408.15252*.

Zhao, S., Li, Y., Yao, X., Nie, W., Xu, P., Yang, J., & Keutzer, K. (2020, October). Emotion-based end-to-end matching between image and music in valence-arousal space. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2945-2954).

Zou, J., Mei, J., Ye, G., Huai, T., Shen, Q., & Dong, D. (2023, October). EMID: An Emotional Aligned Dataset in Audio-Visual Modality. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice* (pp. 41-48).