ReSilence

Retune the Soundscape of future cities through art and science collaboration

HORIZON- 101070278

D3.1

Audio, Visual and Multimodal Analysis Tools

Dissemination level:	Public
Contractual date of delivery:	Month 17, 31 January 2024
Actual date of delivery:	Month 18, 16 February 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:	Report
Approval Status:	Final Draft
Version:	1.1
Number of pages:	31
Filename:	d3.1_resilence_Audio Visual and Multimodal Analysis Tools_v1.1.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.

co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	30/11/2023	ToC and First set of inputs	Antonio Camurri
0.2	03/01/2024	Inputs from partners	All involved partners
0.9	01/02/2024	Submitted for review	Melanie Wald-Fuhrmann
1.0	13/02/2024	Final version after review	All involved Partners
1.1	14/02/2024	Quality control	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.

Abbreviations and Acronyms

ADE	Art-Driven Experiment
AI	Artificial intelligence
cGAN	conditional Generative Adversarial Network
CLIP	Contrastive Language–Image Pre-training
CPU	Central Processing Unit
DNN	Deep Neural Network
FPGA	Field Programmable Gate Arrays
GAN	Generative Adversarial Network
GPIO	General-Purpose Input/Output
GPU	Graphics Processing Unit
HAT	Hardware Attached on Top
IMU	Inertial Measurement Unit
LiPo	Lithium Polymer
MFCC	Mel-frequency cepstral coefficients
NiMH	Nickel–Metal Hydride battery
OS	Operation Systems
PCB	Printed Circuit Board
RAM	Random-Access Memory
SBC	Single-Board Computers
SRAM	Static Random Access Memory (SRAM)
UPS	Uninterruptible Power Supply
USB	Universal Serial Bus
VAD	Voice Activity Detection
VQ-VAE	Vector Quantized Variational Autoencoder

Table of Contents

1 INTRODUCTION

One of the objectives of the ReSilence project is to involve and collaborate with artists 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 scenario, D3.1 focuses on the tools for audio, visual, and multimodal analysis that are developed by ReSilence partners with/for the selected artists. In particular, this deliverable presents a brief overview of such tools at this intermediate stage of the project.

In Section 2 the Methodology is presented via the Research Objectives and the expected results.
In Section 3 briefly presents the relation of WP3 activities to the high-level requirements as described in D6.1 (Pilot use cases and initial requirements and challenges).
Sections 4 focus on using sensor fusion and artistic theories to analyse expressive qualities of human movement and group dynamics through non-verbal signals in urban spaces, emphasizing cohesion and synchronization in shared musical experiences.
Sections 5 reviews software libraries for automated sonic feature analysis in soundscapes and describes field recordings by implementing real-time voice "blurring" and time-triggered recording to distort identity while preserving ambient sounds on low-cost, portable devices.
Section 6 includes AI models for cross-modal learning, synthesizing images from sounds and sounds from images.
In Sections, 4 through 6, information about the equipment used is provided, along with the applied techniques. Also, a short description of goals and requirements is given.
Section 7 concludes this deliverable.

2 METHODOLOGY

This deliverable is part of WP3, focusing on the development and refinement of algorithms and techniques for real-time interactions in the project use cases. The work in work package 3 has one research objective (RO):**RO1_AI based technologies for real-time interaction.**

RO1 focuses on developing AI technologies that enable real-time interaction, enhancing both individual and collective engagement with music and sound, thereby building trust in these technologies among citizens and participants.

The Research Activities (RA) that are being conducted with respect to RO1 are the following:

RA1.1: Multimodal movement analysis and sonification

RA1.2: Audiovisual immersive soundscape analysis

RA1.3: AI-based soundscapes

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;
- Record anthropogenic sounds in real time while conforming with the GDPR through voice activity detection and blurring.
- Cross modal learning aiming at the development of an audio-to-image and image-to-audio synthesis model;

3 RELATION TO USER REQUIREMENTS

The following table accumulates the user requirements that have been developed until now, based on the artists' expression of their needs to tackle the challenges tackled through ReSilence project. The user requirements are taken from the use cases section (D6.1) 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 amount 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

4 MULTIMODAL MOVEMENT ANALYSIS

This section refers to the activities in T3.1. In particular, T3.1.1 focuses on algorithms for real-time estimation of expressive qualities and intent from human full-body movement, as measured from environmental and body worn sensors. We start by identifying the multimodal interaction models and the sensor systems, including cameras, body worn IMUs, mobile devices, and proximity sensors, based on results obtained in previous and running EU projects (ICT DANCE, ICT Wholodance, FET PROACTIVE EnTimeMent). The development of a sensor fusion framework is under development and will be used to extract features from body movement, including mid-level features, e.g. impulsivity, fluidity, lightness, directness, heaviness. The estimated features will be used to estimate expressive content in movement, including affective qualities, interest, and engagement (e.g., hesitation, fragility, directness of movement towards a target) of a person moving through simulated city spaces. Both features learned from data and features inspired by artistic and humanistic theories, such as Rudolf Laban's Theory of Effort are considered. Algorithms will be validated with both existing datasets and with data collected within the use case scenarios.

T3.1.2 faces the problem of measuring features related to non-verbal social signals of users moving in city spaces (real or simulated, e.g. in interactive artistic installations): in particular, the focus is on synchronisation, entrainment, leadership in a group of individuals: the objective is to measure shared joint multisensory experience of music by a group of users, e.g., their non-verbal full-body synchronisation and dialogue by moving in a space. Cohesion is generally considered as the group members' inclinations to forge social bonds, resulting in the group sticking together or the "individual's personal motivations to remain in the group". Entrainment is modelled in its temporal and affective components (Phillips and Keller, 2012). The ongoing work is grounded on previous research results by UNIGE. Techniques such as Recurrent Quantification Analysis, Event Synchronisation, Granger Causality, or Sample Entropy, and other techniques developed by UNIGE (e.g. MECS) are under investigation and will be considered for the scientific experiments and the artistic projects of the selected artist, in particular in the project of Andrea Cera.

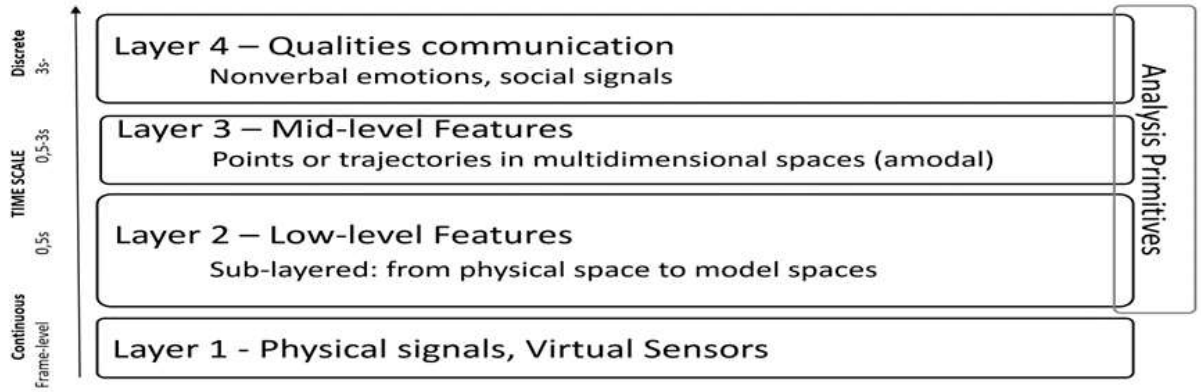

Figure 1: Conceptual framework for multimodal movement analysis

Feature	Type	Relative joint(s)
Kinetic energy	Holistic, Kinematic	All
Contraction index	Holistic, Postural	All
Symmetry	Holistic, Postural	All
Postural Attitudes	Holistic and local, Postural	Trunks joints, head
Periodicity	Holistic and local, Postural	All, shoulders, hands
Direction of movement	Holistic and local, Postural	All, hands, elbows
Kinematics	Local, Kinematic	Hands, elbows, shoulders
Fluidity	Holistic, Kinematic	All
Impulsiveness	Local, Kinematic	Hands
Curvature	Local, Trajectory	Hands, elbows
Smoothness	Local, Trajectory	Hands
Distance from body	Local, Postural	Hands, elbows
Distance from head	Local, Postural	Hands

A preliminary list of Movement Features proposed in the collaboration with the artistic project of Andrea Cera and in other artistic projects. These factors comprehend simple kinematics (i.e., velocities, accelerations) of single joints, spatial occupation-related features computed on group of joints, and more complex features (i.e., movement fluidity, periodicity, symmetries) computed on the whole body.

4.1 Related work

Previous work in this sector includes scientific results and technology developments (in particular software libraries for movement analysis) from previous EU projects, including FP7 ICT Wholodance (wholodance.eu), the H2020 FET PROACTIVE EnTimeMent project (entimement.dibris.unige.it) and the DanzArThe project (Camurri et al. 2016, 2024; Piana et al. 2016; Niewiadomski et al. 2019; Vaessen et al. 2019).

4.2 Hardware equipment

Hardware equipment depends on the specific application and is scalable, from motion capture systems (Qualisys) to wearables (e.g., IMUs) and video cameras. The ongoing activities include the investigation and interaction design of the system architecture to be adopted in joint scientific experiments and artistic applications, in collaboration with a number of the selected artists.

.

4.3 Dataset(s)

A number of datasets is available from UNIGE on annotated full-body movement recordings, based on mocap, video cameras, microphones, IMUs, created in the above-mentioned EU projects and in other collaborations with artists. We plan to record novel multimodal datasets in the context of some of the ReSilence artistic projects: these will be considered also for validation of ongoing scientific work on improvement of the above-mentioned movement features and on the identification of new ones.

4.4 Participants

Participants will include audience members in the concerts, installations, and rehearsals, as well as participants recruited for specific scientific experiments necessary in the trans-disciplinary interaction design process of the artistic projects.

4.5 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.

.

5 SOUND AND SPACE SENSING AND SOUNDSCAPE ANALYSIS

In this section we summarise our analysis of the state of the art of software libraries that could be candidates for the automated analysis of sonic features in soundscapes. They include third party as well as partners' software libraries, and a survey of the literature of available sound datasets, in particular of soundscapes. Details of the dataset adopted in the project by UNIGE in the collaboration with the artistic project and related scientific experiments of the artist Andrea Cera are also provided. Additionally, in the same approach, the support provided to Caroline Clauss by CERTH is described.

5.1 Supporting Caroline Clauss' experiments and project challenges

The soundscape in urban environments/spaces contains acoustic signals that may be either environmental or anthropogenic. For projects involving human activity (speech, anthropogenic noise), like Caroline Clauss', field recording is a sensitive matter being protected under the GDPR. Therefore, the main issue that arises is related to handling human voice activity (speech, hue, timbre, etc.), without eliminating it from the soundscape, during the recording process, and before the data is stored on a respective device. Given the fact that recording the soundscape requires long hour recording files, issues arise with power consumption as well.

Our proposed solution for this challenge is to "blur" the voices to change the characteristics that denote identity, as well as to render speech content incomprehensible. In more detail, the human voice is composed of fundamental frequencies and their corresponding harmonics. The voice frequencies are placed roughly between 20Hz and 20kHz. Yet, this spectral range is not limited to humans, other urban sounds may coexist. Therefore, voice detection, source separation and voice manipulation are necessary.

Moreover, to record in real time while conforming with the GDPR, requires special equipment. Additional challenges arise in implementing this voice-blurring process on a low-cost, low-energy, portable device, given the critical importance of preserving the presence of ambient voices without simply truncating them. That is, presenting a viable and technically feasible approach to safeguarding privacy while maintaining the integrity of the environmental sound data by proposing a solution that allows distortion of the voice in real-time during the recording and detection phase. To save system resources (e.g. storage space, energy consumption), time triggered recording is necessary. Rush hours, or high interest sound activity hours will be defined to set the recording duration and period, on said equipment.

5.1.1 Related work

As far as field recordings are concerned, there have been projects that collect environmental sounds in the context of soundscape ecology, to analyse the biodiversity of an area (Pronk, 2022). These approaches are devoid of the human factor therefore not appropriate for urban environments. To ensure personal privacy during recording in urban spaces, signal processing techniques need to be utilised for concealing the human ID, or the speech content. The first step to achieve the above is by identifying which part of the audio signal contains the human voice signal, as well as by utilising voice activity detection (VAD). For this specific implementation, the direction was to find low-powered VAD implemented algorithms.

.

5.1.1.1 **Voice activity detection and blurring**

Accurate detection is challenging, particularly when the speech signal is corrupted by noise. Many VAD systems have been proposed and widely studied in the past decade. They can be categorised into two major types: software VAD and hardware VAD. In software VAD systems, various features are introduced, such as spectral entropy, hidden Markov models, cepstrum coefficients, and others, which are reviewed and compared in (Alías et al., 2016). A more complicated VAD algorithm based on a deep neural network (DNN) learns features during the training process (Liu et al., 2019). Despite their high functionality, most of these methods require complex computations, making them inappropriate for low-cost and low-power hardware implementation.

For the ReSilence project, low-powered VAD implemented algorithms are the optimal solution since the algorithmic calculations will take place on portable devices. For similar purposes, and more specifically to run via Field Programmable Gate Arrays[1] (FPGA) a speech recognition preprocessor, a serial logistic regression classifier which uses as features a) the frame-energy, b) the maximum absolute signal finite difference and c) the maximum absolute squared signal finite difference over a frame, has been developed (Meoni et al., 2018). After an evaluation with a similar serial VAD approach, it was obvious that the device was able to operate by consuming nearly 75% less energy (0.56mW instead of 2.10mW), even though the occupation area was low. Tests need to be implemented to determine the coverage area according to the recording equipment.

In a recent effort, Faghani et al. (2023) used level-crossing sampling (where the non-speech parts of the signal are eliminated) with a 40 ms moving window with a 30 ms overlap as a feature extraction block (Figure 2). In a simulation environment that includes different noise types and SNR levels, the evaluation results indicated a significant reduction of the sampling rate and power consumption (394.6 nW), maintaining acceptable accuracy levels.

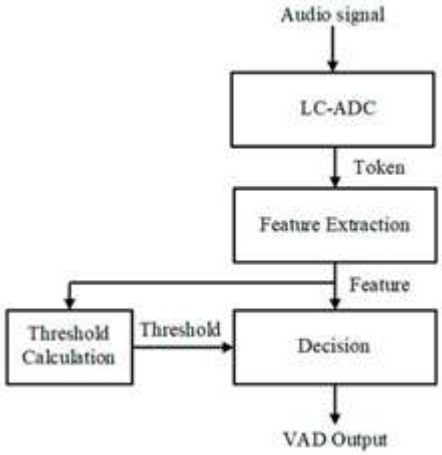

Figure 2: Block diagram of the VAD algorithm proposed by Faghani et al. (2023)

To voice blurring, Cohen-Hadria et al. (2019) proposed a novel approach to voice anonymization in urban sound recordings. The proposed method is inspired by the process of face blurring in images and is the first to achieve the three objectives of speaker de-

[1] https://www.xilinx.com/products/silicon-devices/fpga/what-is-an-fpga.html

identification, content obfuscation, and scene preservation in environmental sound recordings. The method uses a deep U-Net source separation model to extract voices from non-voice content and applies low pass filtering and MFCC inversion to obscure the identity and intelligibility of the extracted voices. Given the source separated and identified voice signal, Kai et al. (2021) propose a framework for lightweight and irreversible speech pseudonymization based on data-driven optimization of cascaded voice modification modules. It utilises both machine learning-based and signal processing-based approaches to pseudonymize speech while maintaining intelligibility and suppressing speaker-specific information incorporating cascaded and superposition-based voice modification modules and employing data-driven optimization of the parameters of the voice modification methods.

5.1.1.2 **Hardware equipment for real time data acquisition**

The equipment necessary for high quality recordings may be expensive or intrusive to the environment/society, a restrictive factor for the researchers. To cope with this issue and in parallel with Open Science rise, engineers have cooperated with other scientific fields to produce hardware and software specialised in their particular needs (Gibb et al., 2019). Popular paradigms of passive acoustic recording devices are the AudioMoth (Hill et al., 2019), Solo (Whytock and Christie, 2017), and microPAM[2]. These devices have good audio recording characteristics, may provide band-pass filtering, and have been implemented mainly in the wild, where the soundscape is lacking anthropogenic sounds.

For more complex tasks, single Board Computers (SBC) like the by far most popular Raspberry Pi have been used, among others, for monitoring the ecosystem behaviour by analysing the soundscape, but also for supporting urban soundscape classification (Florentin and Verlinden, 2017; Arce et al., 2012; Segura-Garcia et al., 2021). Raspberry Pi supports all possible research approaches (interactive, autonomous, long-distance, trigger oriented, etc.), of various input (audio, video, etc.), providing environmental monitoring either in controlled spaces (e.g. laboratory experiments), or at field measurement stations, may process data automatically and in real-time, send notifications, and more (Jolles, 2021). Recordings that require complex preprocessing, naturally rely on the Raspberry Pi computing power.

Other approaches, like the microPAM (Schank and Riesbeck, 2013), utilise the electronic platform Arduino or the Raspberry Pico processor to materialise a low-cost passive acoustic monitor that is also used for various applications[3]. Research projects may deploy other Arduino equivalents, like the USB-based microcontroller development system Teensy (Suitor et al., 2024). In general, the main criterion for equipment selection depends on each project's particular needs (computational, environmental, economic, educational, etc.).

5.1.2 **Hardware equipment**

In the framework of ReSilence, certain needs regarding recording devices have emerged a) according to the artists' vision, b) to conform with the GDPR legislation, or/and c) to utilise the equipment already obtained by the residency artists. Sensors for audio recording like the AudioMoth device, or computers like the Raspberry Pi or Arduino can be combined to attain real-time sound analysis, prior to saving the audio signal collected from urban environments.

[2] https://www.micropam.com/microPAM-T4_hardware.html

[3] Platforms, U. I., & Platforms, C. C. (2018). Content.

Certain devices are being configured to support the data collection experiments for which the artists have the responsibility.

5.1.2.1 **AudioMoth**

AudioMoth is an equipment already purchased by Caroline Clauss, aiming to record sound at urban spaces. For its utilisation, the device has undergone comprehensive study, and alternative solutions have been devised.

AudioMoth is a device that has the ability to record full-spectrum (recording time, frequency, and amplitude) and a wide frequency range (from audible frequencies well into ultrasonic). Its low-cost, frequency range, and digital noise isolation circuitry render it a good choice as an audio sensor (Hill et al., 2019). It can collect uncompressed audio which is stored to a microSD card at rates from 8,000 to 384,000 samples per second. Under MIT licence, several applications have been developed that allow time or frequency triggered band-pass filtering (all, high, low, and band -pass).

The device's open-source firmware is allowing developers to build project- specific applications, yet the device's processor (EFM32 Gecko[4]), although high performance for simple tasks like band-pass filtering, cannot support complex real-time acoustic signal analysis or transformations, nor is it able to perform computationally heavier tasks (e.g. supporting AI). Additionally, the external static random-access memory (SRAM) is used as a buffer in low energy mode to avoid waking up the processor, aiming to achieve battery power conservation.

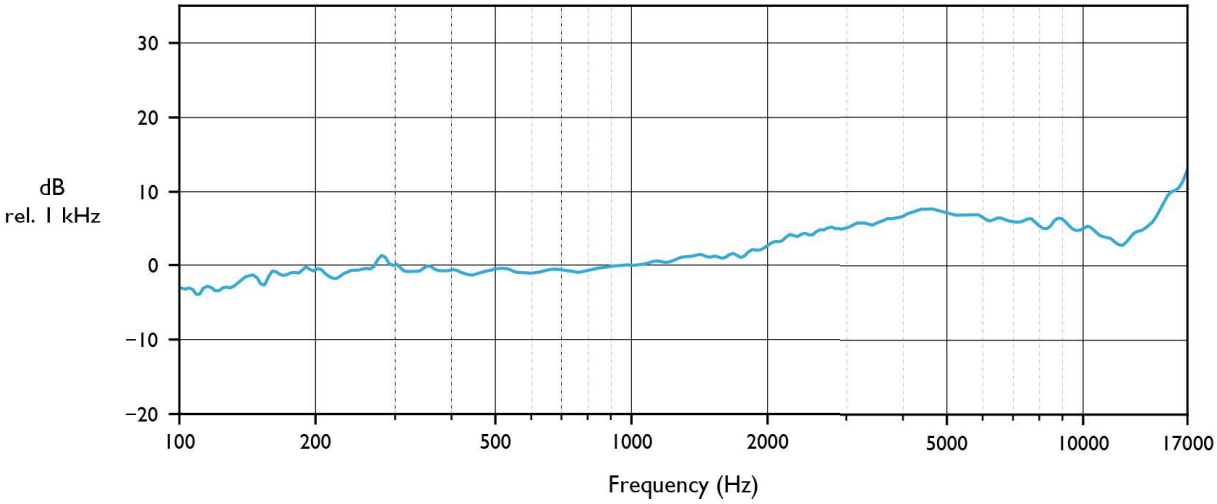

Figure 3: AudioMoth frequency response without protective case (Lapp, 2021)

AudioMoth has the potential to be converted into a full-spectrum USB microphone with the characteristics depicted at Figure 3, by installing the AudioMoth USB Microphone firmware on the device. Thus, it can be combined with more powerful computers or processors (e.g. Raspberry Pi, Arduino, laptops) to collect and process audio signals of high quality. Tests are to be carried out to examine the device's potential.

5.1.2.2 **Raspberry Pi**

The Raspberry Pi is a low-cost SBC, counting several generations of releases. Each model features a system-on-a-chip that consists of an integrated CPU (central processing unit) and

[4] https://www.silabs.com/mcu/32-bit-microcontrollers/efm32-gecko

on-chip graphics processing unit (GPU) and on-board memory. All models also have a port to connect a dedicated camera, as well as an array of general-purpose input/output (GPIO) pins that can be used to communicate with a wide range of electronics. So, it has the functionalities of a standard computer: connect external devices (mouse, keyboard, screen, microphone etc), it supports various Operation Systems (OS) (Linux Desktop environment, Windows 10 IoT, and Android), internet connection (both wireless and through ethernet), and it can be programmed to run scripts autonomously using a range of programming languages.

In the framework of ReSilence, the Raspberry Pi 4B model will be used for audio signal recording, since it is a powerful computer that can support signal processing. This model has various variants that allow the users to select different RAM sizes, up to 8GB RAM (which will be used in Caroline Clauss' project). The AudioMoth microphone is utilised as an audio input device by connecting it to a USB port. Raspberry Pi will provide the computations required for audio signal analysis transformation, to ensure that the recorded audio data complies with the GDPR in real time, ensuring privacy in urban spaces.

Limitations arise regarding the portability of the Raspberry Pi since it requires a power input of 5V DC and 3A. Special expansion boards that connect to the GPIO pins, called Hardware attached on top (HAT), can provide autonomy (power management), or even various functionalities (e.g high-quality audio recording and more). A search regarding possible solutions on power autonomy has provided the following choices for the artists' equipment set-up.

The artists who will purchase the equipment select the power supply choice according to their budget, needs and preferences. Specific description and guidance regarding the cons and pros of each selection has been provided.

5.1.2.3 **UPSs HAT**

Uninterruptible power supply (UPS) are devices that render a computer running for a short time when incoming power is interrupted. There are various UPS products for Raspberry Pi on the market that use rechargeable batteries to power supply it, while at the same time they include integrated microcontrollers that can manage a soft shut down functionality for the Raspberry Pi when the battery is depleted to low charge levels (below 5V). The more robust UPS HATs with a) guaranteed quality circuitry, b) scripts or apps that do not need debugging, and c) available user manuals are:

1. The PiJuice HAT[5]: a portable power platform provided by the Raspberry Pi Foundation. It is an expensive solution that provides a rather small rechargeable battery (1820 mAh) that can be upgraded by purchasing greater capacity batteries that are also quite expensive. It can be connected to renewable power sources.
2. The Pi Sugar S Plus[6]: a reasonably priced alternative with a 5000mAh battery on board.

5.1.2.4 **Power Bank**

Various experiments have been contacted from bloggers[7] and vloggers[8] to power supply the

[5] https://uk.pi-supply.com/products/pijuice-standard

[6] https://www.pisugar.com/#

[7] https://www.powerbankexpert.com/best-raspberry-pi-power-bank/ Ads

[8] "A Great Source of Battery Power for Raspberry Pi! Among other things...(Tech Review)", https://www.youtube.com/watch?v=F4bA_CpVcU8

Raspberry Pi with a simple approach that does not require complex set-up from the user (both in software or hardware), or high-cost equipment. Power banks of high quality that provide appropriate output (5V, 3A for Raspberry Pi 4B) without interruptions seem to be able to support the device. It is easy to connect it (via usb port), for people who are not technologically adept. This approach is not sophisticated, it requires constant attention during the operation of the device, to observe whether the battery depletes below 5V (to avoid hard shut down that may damage the device). Also, power banks with overload protection are optimal to guarantee the device's safety. These characteristics, together with the capacity size that must be high enough to support hours of continuous audio recording, raise the purchase value of an appropriate power bank. The capacity of the power bank is also related to its dimensional size and weight, rendering the set-up more intrusive and exposed to the environment.

5.1.2.5 **Custom made power supply circuits**

Another economical approach is soldering battery boxes (either for rechargeable AA NiMH) or Lithium Polymer (LiPo) batteries, together with a DC/DC converter and a charge controller. The DC/DC converter will convert the 3.7V battery voltage to 5V to guarantee the power supply threshold of Raspberry Pi 4B, while the battery charge controller will regulate the incoming current and voltage to the batteries, to prevent overcharging. There are available PCBs on the market to guarantee a prolonged battery lifespan and the Raspberry Pi safety. This solution was rejected due to unforeseen conditions (weather, environmental and social interactions) that may damage the equipment package, since it is the least robust among the solutions.

5.1.3 **Datasets**

In the context of creating a dataset for recorded soundscapes, existing datasets play a crucial role in labelling and analysing audio samples. These datasets, with well-defined labels, serve as valuable references for identifying common acoustic patterns and categorising different urban sounds. Leveraging this pre-existing knowledge enhances the efficiency of labelling processes and contributes to a better understanding of the acoustic landscape. In detail, for the purpose of sound classification, four datasets are currently taken into consideration.

5.1.3.1 **AudioSet**

AudioSet dataset is a large-scale collection of human-labelled 10-second sound clips drawn from YouTube videos. For the collection of data, human annotators who verified the presence of sounds they heard within YouTube segments, were employed. The nomination of segments for annotation relied on YouTube metadata and content-based search.

5.1.3.2 **UrbanSound**

This dataset contains 1302 labelled recordings of urban sounds, each one labelled with the start and end times of sound events from 10 classes drawn from the urban sound taxonomy. Each recording may contain multiple sound events, but for each file only events from a single class are labelled. All recordings were obtained from www.freesound.org.

5.1.3.3 **ESC-50**

ESC-50 is a dataset for environmental sound classification. It consists of 5-second-long recordings organised into 50 semantic classes loosely arranged into 5 major categories: animals, natural soundscapes and water sounds, human-non-speech sounds,

.

interior/domestic sounds, exterior/urban noises. Clips in this dataset have been manually extracted from public field recordings gathered by Freesound.org.

5.1.3.4 **Emo-Soundscapes**

Emo-Soundscapes allows studying soundscape emotion recognition and how the mixing of various soundscape recordings influences their perceived emotion. It contains 1213 6-second Creative Commons licensed audio clips, based on the curation of 600 soundscape recordings in Freesound.org and 613 mixed audio clips from combinations. For the collection of the ground truth annotations of perceived emotion in 1213 soundscape recordings, a crowdsourced listening experiment was used.

5.1.4 Future considerations

Test recordings with the AudioMoth + Raspberry Pi 4B equipment will assist in optimising the algorithmic variables, delimiting the distance from the sound source, revealing the optimal recording duration and period, and in discovering the general capabilities of the hardware - software. Obstructions coming from circuit cases or other components that may alter the soundscape will also be considered.

5.2 Conceptual Frameworks for Interactive Sonification as used in the artistic project of Andrea Cera

The model of interactive sonification of human movement qualities developed by UNIGE in collaboration with Andrea Cera is based on cross-modal mappings of movement qualities at different levels of abstraction and temporal scales, elaborated in previous EU projects. The main criteria and guidelines for our model are described in the following:

1. Multiple Layers Cross-Modal Correspondence between movement and sound (e.g., Spence 2011): this is the main component of the sound architecture, grounded on previous research (Niewiadomski et al. 2019; Camurri et al. 2016; Piana et al. 2016), identifying a four layers model (see above) 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 (e.g. Botteldooren and De Coensel 2009).

3. Slowness and continuity in dynamic curves: movement fluidity is a mid-level quality (see conceptual framework depicted above, and 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, and 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 semiconsciously 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. 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 perceivable 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, the sonification of movements of the avatar and of the user have a clearly separated “signature”.

5.3 Software libraries for the automated analysis of soundscapes for the artistic project of Andrea Cera

In the first phase of the ReSilence project, several software libraries were analysed and tested in the framework of the artistic project of Andrea Cera, in particular for the automated analysis of perceived *annoyance* and *intrusiveness* of a soundscape. In the following we provide brief details of these tests.

.

An in-depth analysis of the scientific literature on sound annoyance and intrusiveness suggested a subset of the algorithms: in particular, the analysis of literature suggests algorithms that use the following features: Loudness, Roughness, Sharpness, Skewness, Spectral Centroid, classifySound (events detection).

Audio Toolbox, a software library from Matlab, has been utilised. Other software libraries already utilised in previous work (Niewiadomski et al. 2019) that are adopted in ReSilence include the following:

Max/MSP platform for real-time audio processing (in particular for the real time interactive installations of the artistic project of Andrea Cera): ZSA.DESCRIPTORS (software library by Mikhail Malt and Emmanuel Jourdan, IRCAM), in particular algorithms on Spectral Centroid, Flatness, Kurtosis, Roughness, Spread, Flux, Rolloff, Skewness, Slope.

SOUND DESIGN TOOLKIT (software library by IUAV and IRCAM - Davide Rocchesso et al.): algorithms for the analysis of Spectral Centroid, Spread, Skewness, Kurtosis, Flatness, Flux.

Sonic Visualiser (Queen Mary University of London): libxtract (Jamie Bullock).

YAAFE: https://yaafe.sourceforge.net/ (Python / MATLAB)

MOSQITO: https://github.com/Eomys/MoSQITo (Python)

Audio Signal Processing on Raspberry Pi:

CMSIS DSP Software Library[9]: a suite of common signal processing functions for use on Cortex-M and Cortex-A processor-based devices. Useful on programming the Raspberry Pi 4B (that has a Quad core Cortex-A72 processor).

PyAudio Library[10]: Necessary to perform recordings (in this case with AudioMoth as a microphone) through a Raspberry Pi.

Next steps include the definition of prototypes and testbeds for the artistic project of Andrea Cera and the creation of the datasets for the scientific experiments with UNIGE and University of Maastricht.

Furthermore, UNIGE, UM and other ReSilence partners are conducting periodic meetings with the artists Lea Sikau and Loukia Tsafoulia. These meetings constitute the early stage of the interaction design process of these trans-disciplinary projects.

[9] https://arm-software.github.io/CMSIS_5/DSP/html/index.html

[10] https://pypi.org/project/PyAudio /

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. The primary goal is to generate visual representations from captured sound signals and, conversely, generate audio from images. 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 audio-to-image and image-to-audio models. The derived tools will be aligned with the artistic-driven approach of the ReSilence project, based on the needs of each collaborating artist.

6.1 Related work

During months 3-17, a literature review has taken place both for AI-based image and audio generation methods. A brief description of the most prominent state-of-the-art works follows.

6.1.1 Audio-to-image generation

Cross-modal audio-visual perception explores how humans perceive and make sense of stimuli when presented with both sound and visual cues simultaneously. It has been a long-lasting topic in psychology and neuroscience, and various studies have discovered strong correlations in human perception of auditory and visual stimuli. In Chen et al. (2017), the authors attempt to address this challenge of cross-modal generation by leveraging the power of deep generative adversarial training. Specifically, they use conditional Generative Adversarial Networks (cGANs) to achieve cross-modal audio-visual generation of musical performances. Different encoding methods are explored for audio and visual signals, and two scenarios are proposed: instrument-oriented generation and pose-oriented generation.

In Wan et al. (2019), the authors introduce a novel method in which images are generated conditioned on sounds. Based on the SoundNet dataset (Aytar et al., 2016), they utilise image and sound classification results to build a relatively cleaner image-sound paired dataset. By applying different methods to the proposed generative model, the model can generate images with better quality in terms of both subjective and objective evaluations.

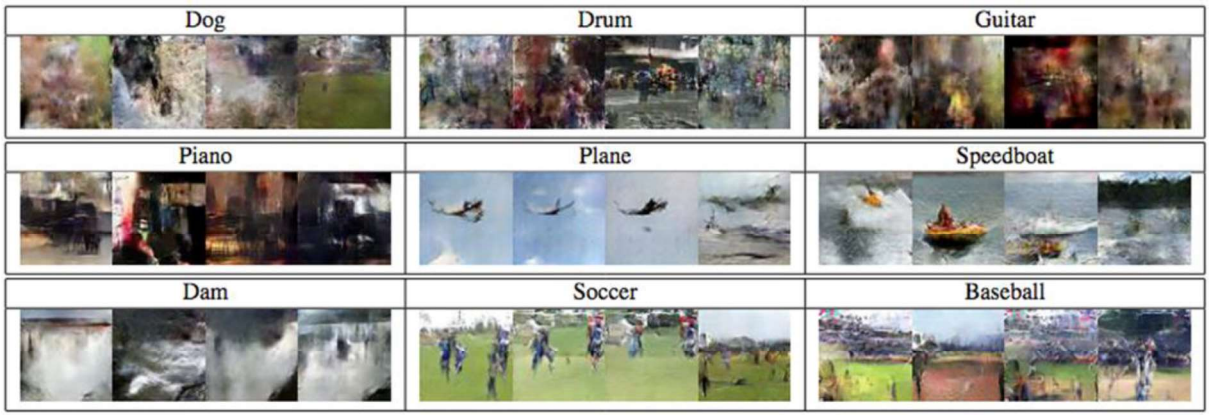

Figure 4: Samples of audio-to-image generation from Wan et al. (2019)

.

Another work, ImageBind (Wan et al., 2023), is the first AI model that can bind data from six modalities (image/video, audio, text, depth, heatmap and IMU) at once. To achieve this, it leverages web-scale (image, text) paired data and combines it with naturally occurring paired data such as (video, audio), (image, depth) etc. to learn a single joint embedding space. This allows ImageBind to implicitly align the text embeddings to other modalities such as audio, depth etc., enabling zero-shot recognition capabilities on that modality without explicit semantic or textual pairing. This method can be applied to a variety of different modalities and tasks with little training.

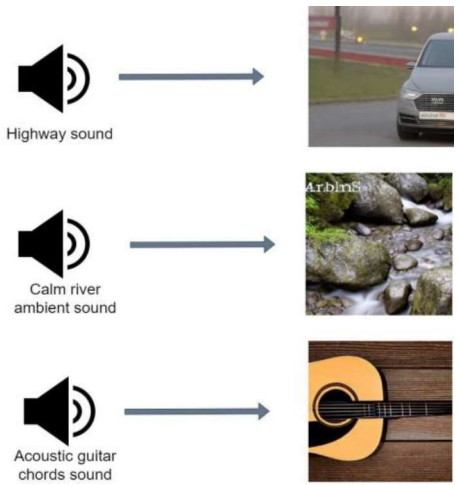

Figure 5: Examples of audio-to-image generation using ImageBind

6.1.2 **Image-to-audio generation**

IM2WAV (Sheffer and Adi, 2023) is an image-guided open-domain audio generation system. Given an input image or a sequence of images, it generates a semantically relevant sound. This method is based on two transformer language models that operate over a hierarchical discrete audio representation obtained from a VQ-VAE (Vector Quantized Variational Autoencoder)-based model. At first, a low-level audio representation is produced using a language model. Then, the audio tokens are upsampled using an additional language model to generate a high-fidelity audio sample. This method uses the rich semantics of a pre-trained CLIP (Contrastive Language–Image Pre-training) (Radford et al., 2021) model embedding as a visual representation to condition the language model. In addition, to steer the generation process towards the conditioning image, the classifier-free guidance method is applied.

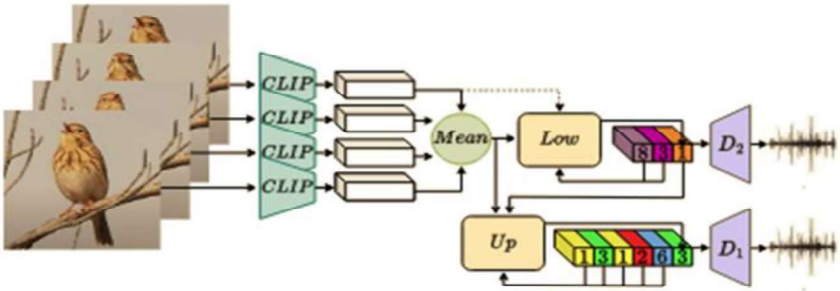

Figure 6: High-level description of theIM2WAV architecture (Sheffer and Adi, 2023)

6.1.3 **Audio-reactive image/video generation**

Audio-reactive visuals generation refers to the process of creating dynamic visual content that synchronises and responds in real-time to an audio input, such as music or sound. It involves using algorithms and techniques to generate visual patterns, colours, or animations that are influenced by the characteristics of the accompanying audio, like rhythm, frequency, or volume. This concept is commonly employed in music visualisation, live performances, and interactive art installations to create captivating and synchronised visual experiences. Some state-of-the-art works are presented in the next paragraphs.

According to a method proposed for creating audio-reactive visuals based on StyleGANs[11] (Karras et al., 2019), the audio-reactive visuals that are produced follow the magnitude changes in the frequency band of the audio input to traverse the latent space of GANs. As the latent space is smooth and interpolatable, by concatenating the generated images a smooth video clip can be formed, which reacts to the audio clip. StyleGan is used as it maps a latent vector to several styles that control the coarse-to-fine structures of the image, which makes it intuitive to map to when there are multiple features. Nsynth (Oord et al., 2016) is also explored to extract features from the audio clip, and GAN steerability (Jahanian et al., 2019) is applied to learn specific walks in latent space that correspond to effects such as zooming and rotation.

StyleGAN allows for fine-grained control of image generation through its hierarchy of latent and noise inserts. Using musical information like onset envelopes and chromagrams, latent vectors and noise maps can be generated to create interpolation videos that react to music. (Brouwer, 2020) introduces techniques that create convincing audio-reactive videos by combining multiple interpolations that react to individual musical features. The amount of musical information that can be visually recognized is maximised by mapping different musical features to different parts of the latent and noise hierarchy.

Deepsing (Passalis and Doropoulos, 2021) is a deep learning method for performing attributed-based music-to-image translation. This method is applied for synthesising visual stories according to the sentiment expressed by songs. The generated images aim to induce the same feelings to the viewers, as the original song does, reinforcing the primary aim of music, i.e., communicating feelings. The process of music-to-image translation poses unique challenges, mainly due to the unstable mapping between the different modalities involved in this process. Deepsing employs a trainable cross-modal translation method to overcome this limitation, leading to a deep learning method for generating sentiment-aware visual stories.

6.2 **Datasets**

During months 3-17, we conducted research on available datasets to be used for our experiments for audio-to-image and image-to-audio generation. We mainly researched datasets that contain pairs of images and sounds or videos, as well as datasets that associate emotions with visual content and visual datasets used for style transfer. The most intriguing datasets are briefly presented in this subsection.

[11] https://hanhung.github.io/Creating-Audio-Reactive-Visuals-With-StyleGAN/

6.2.1 Image-sound pairs datasets

6.2.1.1 SoundNet

The SoundNet dataset[12] contains videos crawled from the Web used to train the SoundNet classification model to classify sounds.

6.2.1.2 ImageHear

ImageHear is a dataset proposed in (Sheffer and Adi, 2023) that contains pairs of images and their corresponding sounds to train audio-to-image or image-to-audio generation models.

6.2.1.3 Sub-URMP

The Sub-URMP (Chen et al., 2017) dataset contains image and audio files cut from the original URMP (Li et al., 2016) videos. A sliding window method is used to obtain the samples. The size of the sliding window is 0.5 seconds, and the stride is 0.1 seconds. The first frame of each video chunk is used to represent the visual content of the sliding window. The audio files are in WAVE format with a sampling rate of 44 KHz and bit depth of 16 bits, stereo channel. The image files are 1080P (1080x1920). This dataset is used for deep cross-modal audio-visual generation.

6.2.1.4 TAU Urban Audio-Visual Scenes

TAU Urban Audio-Visual Scenes (Wang et al., 2021) is a location classification dataset that consists of 12292 videos and sound clips categorised into 10 classes: airport, metro station, public square, shopping mall, street pedestrian, street traffic, travelling by bus, travelling by metro, travelling by train, and urban park.

6.2.1.5 ADVANCE

The ADVANCE (AuDio Visual Aerial sceNe reCognition datasEt) (Hu et al., 2020) consists of 5075 paired images and sound clips categorised into 13 classes: airport, beach, bridge, farmland, forest, grassland, harbour, lake, residential area, orchard, sparse shrub land, sports land, and train station.

6.2.2 Style transfer/visual emotion datasets

6.2.2.1 WikiArt

The WikiArt dataset contains 81,444 pieces of visual art from various artists, taken from WikiArt.org, along with class labels for each image referring to “artist”, “genre”, “style”.

6.2.2.2 WikiArt Emotions

The WikiArt Emotions dataset (Mohammad and Kiritchenko, 2018) comprises 4,105 artworks, predominantly paintings, annotated to capture the emotions they express. These emotional annotations are obtained through crowdsourcing, categorising the art into one or more of twenty emotion categories, including a neutral category.

[12] http://soundnet.csail.mit.edu/

6.3 Future considerations

Our next steps involve adapting the existing methods to the project's requirements using the datasets that are currently available and any datasets that may emerge in the context of ReSilence. For this reason, the ReSilence partners are collaborating with the artists (currently Caroline Claus and Loukia Tsafoulia) who are interested in AI-based audio and image generation to design the corresponding tools that will meet the needs of their artistic creations.

7 CONCLUSIONS

The ongoing activities in the ReSilence project are demonstrating the trans-disciplinarity involving scientific and technological developments offered by ReSilence partners and the artistic projects by the selected artists. In the next phase of the project, a number of artistic demonstrations, scientific and artistic experiments, showcases, and datasets are evolving and will be demonstrated in public events and scientific papers.

8 REFERENCES

Alías, F., Socoró, J. C., & Sevillano, X. 2016. "*A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds*", *Applied Sciences*, *6*(5), 143.

Arce, P., Salvo, D., Piñero, G., & Gonzalez, A. 2021. *"FIWARE based low-cost wireless acoustic sensor network for monitoring and classification of urban soundscape"*, Computer Networks, 196, 108199.

Aytar, Y., Vondrick, C., & Torralba, A. 2016. "*Soundnet: Learning sound representations from unlabeled video",* Advances in neural information processing systems, 29.

Botteldooren, D. and De Coensel, B. 2009. "*Informational masking and attention focussing on environmental sound*", NAG/DAGA Proceedings, p.399-402.

Brouwer, H. 2020, December. "*Audio-reactive latent interpolations with stylegan*", In NeurIPS 2020 Workshop on Machine Learning for Creativity and Design, vol. 3.

Camurri A, Seminerio E, Morganti W, Canepa C, Ferrari N, Ghisio S, Cera A, Coletta P, Barbagelata M, Puleo G, Nolasco I, Costantini C, Senesi B and 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*", Front. Comput. Sci. 5:1290589. doi: 10.3389/fcomp.2023.1290589.

Camurri, A., Volpe, G., Piana, S., Mancini, M., Niewiadomski, R., Ferrari, N., Canepa, C. 2016. "*The dancer in the eye: towards a multi-layered computational framework of qualities in movement",* Proceedings of the 3rd International Symposium on Movement and Computing MOCO2016, p. 1-7.

Chen, L., Srivastava, S., Duan, Z., & Xu, C. 2017, October. "*Deep cross-modal audio-visual generation",* In Proceedings of the on Thematic Workshops of ACM Multimedia 2017 (pp. 349-357).

Cohen-Hadria, A., Cartwright, M., McFee, B., & Bello, J. P. 2019, October. "*Voice anonymization in urban sound recordings",* In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.

Faghani, M., Rezaee-Dehsorkh, H., Ravanshad, N., & Aminzadeh, H. 2023. "*Ultra-Low-Power Voice Activity Detection System Using Level-Crossing Sampling"*, Electronics, 12(4), 795.

Florentin, J., & Verlinden, O. 2017, July. "*Autonomous wildlife soundscape recordingstation using Raspberry Pi*", In International Congress on Sound and Vibration-ICSV (Vol. 24).

Gibb, R., Browning, E., Glover-Kapfer, P., & Jones, K. E. 2019. "*Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring",* Methods in Ecology and Evolution, 10(2), 169-185.

.

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

Hu, D., Li, X., Mou, L., Jin, P., Chen, D., Jing, L., ... & Dou, D. 2020. "*Cross-task transfer for geotagged audiovisual aerial scene recognition"*, In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16 (pp. 68-84). Springer International Publishing.

Jahanian, A., Chai, L., & Isola, P. 2019. "*On the" steerability" of generative adversarial networks",* arXiv preprint arXiv:1907.07171.

Jolles, J. W. 2021. "*Broad-scale applications of the Raspberry Pi: A review and guide for biologists*", Methods in Ecology and Evolution, 12(9), 1562-1579.

H. Kai, S. Takamichi, S. Shiota and H. Kiya, *"Lightweight Voice Anonymization Based on Data-Driven Optimization of Cascaded Voice Modification Modules",* 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 2021, pp. 560-566, doi: 10.1109/SLT48900.2021.9383535.

Karras, T., Laine, S., & Aila, T. 2019. "*A style-based generator architecture for generative adversarial networks*", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

Lapp, S. 2021. "*AudioMoth Performance Testing: A quantitative report of audio recording quality for the AudioMoth",* GitHub repository: https://github.com/kitzeslab/audiomoth-performance.

Li, B., Liu, X., Dinesh, K., Duan, Z., & Sharma, G. 2016. "*Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications*", IEEE Trans. Multimedia, submitted. Available: https://arxiv. org/abs/1612.08727.

Liu, B.; Wang, Z.; Guo, S.; Yu, H.; Gong, Y.; Yang, J.; Shi, L. 2019. "*An Energy-Efficient Voice Activity Detector Using Deep Neural Networks and Approximate Computing*", Microelectronics J., 87, 12–21.

Meoni, G., Pilato, L., & Fanucci, L. 2018, July. "*A low power voice activity detector for portable applications*", In 2018 14th conference on Ph. D. research in microelectronics and electronics (PRIME) (pp. 41-44). IEEE.

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, vol 13, p,191-203.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. 2016. "*Wavenet: A generative model for raw audio",* arXiv preprint arXiv:1609.03499.

Passalis, N., & Doropoulos, S. 2021. "*deepsing: Generating sentiment-aware visual stories*

using cross-modal music translation", Expert Systems with Applications, 164, 114059.

Piana, S., Alborno, P., Niewiadomski, R., Mancini, M., Volpe, G., & Camurri, A. 2016. "*Movement fluidity analysis based on performance and perception",* Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, p. 1629-1636.

Pronk, K. 2022. Using the AudioMoth-a novel passive acoustic monitoring technology-to monitor bat diversity in a rewilded landscape.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. 2021, July. *"Learning transferable visual models from natural language supervision"*, In International conference on machine learning (pp. 8748-8763). PMLR.

Saif Mohammad and Svetlana Kiritchenko. 2018. *"WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art"*, In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Schank, R. C., & Riesbeck, C. K. 2013. *"Micro PAM"*, In Inside Computer Understanding (pp. 180-196). Psychology Press.

Segura-Garcia, J., Calero, J. M. A., Pastor-Aparicio, A., Marco-Alaez, R., Felici-Castell, S., & Wang, Q. 2021. *"5G IoT system for real-time psycho-acoustic soundscape monitoring in smart cities with dynamic computational offloading to the edge"*, IEEE Internet of Things Journal, vol.8 (15), 12467-12475.

Sheffer, R., & Adi, Y. 2023, June. *"I hear your true colors: Image guided audio generation"*, In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Spence, C. 2011. *"Crossmodal correspondences: A tutorial review",* Attention, Perception, & Psychophysics, 73, 971-995.

Vaessen, M., Abassi, E., Mancini, M., Camurri, A., de Gelder, B. 2019. "*Computational Feature Analysis of Body Movements Reveals Hierarchical Brain Organization",* Cerebral Cortex, vol 29 (8), p. 3551–3560.

Wan, C. H., Chuang, S. P., & Lee, H. Y. 2019, May. *"Towards audio to scene image synthesis using generative adversarial network"*, In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 496-500). IEEE.

Wang, S., Mesaros, A., Heittola, T., & Virtanen, T. 2021, June. *"A curated dataset of urban scenes for audio-visual scene analysis"*, In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 626-630). IEEE.

Whytock, R., & Christie, J. 2017. *"Solo: An open source, customizable and inexpensive audio recorder for bioacoustic research"*, Methods in Ecology and Evolution, vol. 8 (3), 308-312.

.