POLITECNICO DI TORINO

Master's Degree in Cinema and Media Engineering

Master's Degree Thesis

Recording of ecological audiovisual scenes for the Audio Space Lab of the Politecnico di Torino

Supervisors

Prof. Arianna ASTOLFI

Prof. Marco Carlo MASOERO

Candidate

Ignazio LIGANI

April 2023

Alla mia famiglia, al mio amico Paolo e a me stesso.
A quest'ultimo auguro di trovare, prima o poi, la pace che sta cercando da anni.

Abstract

Nowadays, a large portion of the elder population suffers from hearing loss, negatively impacting life quality. Therefore, hearing-impaired older adults are often fitted with hearing aids (HAs). However, many complain of insufficient hearing support, especially in complex listening environments. One cause can be found in the standard audiometric tests used to assess HAs performance, which do not account for the actual conditions in which the listening mechanism is activated in real life, as the complexity of the acoustical scenarios and the influence of other senses like the visual component. Thus, to perform ecological listening tests, i.e., tests involving everyday-life conditions, recently, Virtual Reality (VR) systems have been used to reproduce listening tests inside virtual sound environments, representing most frequently attended environments (FAEs) with multiple varying noise sources (e.g. cafes, stations, supermarkets). To support the development of ecological tests, different databases have been published comprising collections of audiovisual simulations or spatial audio recordings of real-life acoustical environments. However, none of them provide in-field recordings of both visual and auditory background, necessary to have a full representation of real-world conditions. Hence, the goal of this thesis was to acquire spatial audiovisual recordings that achieve a high degree of visual and auditory realism and allow to properly test listeners with different noise sources coming from specific surrounding directions. The created scenes are then meant to be played inside the Audio Space Lab room of the Polytechnic of Turin, which hosts a 360°-audiovisual reproduction system composed of a spherical cup of 16-loudspeakers (LSs), able to reproduce audio tracks up to the 3rd-order ambisonics (3OA) encoding, synced with the Oculus Quest 2 VR headset, used to stream the visual content. The main criteria found in the literature that define environmental complexity were considered to select the FAE scenes to be recorded; that is, noise type, position, and distance of target- and noise-sources. Then, to acquire spatial audio recordings that can be played on any reproduction system, the 3OA encoding was chosen, which generically decomposes the sampled sound field into 16 spherical harmonics that can be easily mapped onto any LSs array. After identifying all the possible environments, two of them were selected to start acquiring the audiovisual scenarios: a conference room and a classroom. To allow the auralization of speech audiometry tests inside these environments combined with different kinds of noises and directions, multiple Room Impulse Responses (RIRs) were recorded through the Zylia ZM-1 19-capsules microphone, each for a different configuration of listeners-to-target and -to-noise positions. Similarly, multiple 6K stereoscopic 360° videos sampling the visual scenes were taken using the Insta360 Pro camera. Post-production procedures for both audio and video

tracks were then implemented to erase the unwanted equipment inside the visual scenes and to acoustically compose the different scenes through the convolution of the RIRs with anechoic target and noise signals. As future works, tests on a pool of normal-hearing subjects are planned to validate the created audiovisual scenes. Moreover, further scenes will be recorded also involving the synchronization between sound and lips movement, at first excluded in the current work.

Table of Contents

1 Introduction 1

2 State of the art 4

2.1 Fundamentals 4

2.1.1 Definitions in acoustical field 4

2.1.2 Room Acoustics Parameters 6

2.2 Literature review of common scenes for hearing research 8

2.3 Scenes Databeses 12

2.3.1 ARTE: Ambisonic Recordings of Typical Environments . . . 12

2.3.2 Motus dataset 15

2.3.3 EigenScape 18

2.3.4 Auditory-visual scenes for hearing research 19

2.4 Speech intelligibility: influencing factors in complex environments . 20

2.5 Summary of the environments found 25

2.6 Ambisonics soundfield encoding technique 25

2.6.1 Main concepts & Spherical harmonics 25

2.6.2 Ambisonics orders & formats 28

2.6.3 First-order Ambisonics: coding and decoding 29

2.6.4 Introduction to Higher Order Ambisonics 31

3 Ecological audiovisual scenes realization 33

3.1 Scenes selection 34

3.1.1 Conference Hall 36

3.1.2 Classroom 36

3.2 Equipment for the scenes recording 39

3.3 In-field measurements procedure 44

3.4 Room Acoustic Characterization 47

3.5 Video post-processing procedure 48

4 Applications: scene reproduction inside the Audio Space Lab 71

4.1 Audio Space Lab Structure & Reproduction System 71

4.2 Creation of a scene for speech-in-noise tests 74

5 Conclusions and future perspective 75

Bibliography 79

Chapter 1

Introduction

To contextualize the work done in the realization of this thesis, it's first necessary to make a short presentation of the aim of the project where it will be inserted.

The main objective is to reduce a problem that impacts the life quality of a large part of the elderly population. Indeed, about 25% of people over 60 suffer from hearing problems of various types and intensities [1]. Many of these problems can be solved with the help of hearing aids (HAs), but 83% [2] of the people who could benefit from them do not use them, and 20% of those who own them do not use them [3]. This is because, despite the moderate benefits reported about using HAs in simple situations such as dialogues between two people in quiet environments, there are limited advantages in more complex scenarios.

One of the main reasons for this lack of efficacy lies in the oversimplified audiometric tests used in traditional clinical practice to assess the degree of hearing-impairment. These tests are based on the perception of pure tones and the intelligibility of speech measured by listening to a few words or phrases in quiet or stationary noise conditions [4]. Such a simplified test model doesn't consider the more complex listening situations in which hearing-impaired people can regularly find themselves in their daily life. Furthermore, the non-auditory environmental signals that characterize the perception of the external world, in particular the visual ones, are not taken into account.

The idea to fight against this problem consists in prototyping new standards of audiometric tests to be performed in immersive audiovisual virtual environments, to both diagnose hearing loss and test HAs. These tests will allow evaluation of the hearing sensitivity of patients in virtualized environments inspired by real-life scenarios, compensating for the limitations of traditional audiometric tests.

One of the suitable audio formats for the ecological representation of a spatial sound environment is Ambisonics. It consists in a multi-channel signal in which the sound field around the listener is represented keeping consistent the directionality of the sound sources. The main advantage of this technique is the loudspeaker-independent sound field representation: an Ambisonics recording can be decoded and adapted for any sound reproduction system[5].

Currently, an installation has already been set up at the Polytechnic of Turin in a room called Audio Space Lab (ASL) which will be used for performing the ecological listening testing. The ASL is a sound-treated room that contains a system of 16 loudspeakers plus 2 subwoofers driven by a PC through a 32-channel sound card. The 16 loudspeakers are spherically disposed around the listening positions in the center of the room and are synced with the Meta Quest 2 headset for virtual reality, used to reproduce the desired visual stimuli. The specific loudspeaker arrangement allows recreating in the sweet spot immersive spatialized realistic sound environments reproduced leveraging the 3$^{\text{rd}}$-order ambisonics technique (3OA).

The aim of this thesis is to acquire spatial immersive AudioVisual (AV) in-field recordings of everyday-life scenes where speech intelligibility tests can be auralized and that achieve a high degree of visual and auditory realism while allowing to properly test listeners with different noise sources coming from specific surrounding directions. Specifically, the required characteristics of the database to be developed are:

- acoustical and visual realism;
- selection of the most frequently attended environments by Hearing-Impaired Older Adults (HIOA) with adverse acoustical conditions;
- in-field Room Impulse Response (RIR) recordings in different locations inside each environment to emulate different scenes, namely different spatial configurations for target speech, interfering noise, and listener;
- spatial audio recordings encoded in a loudspeakers-independent format to enable the reproduction of the provided AV scenes on any AV playback system;
- 360° video shootings of the scenes;
- thoroughly acoustical characterization of each environment, performed via in-field measurements of standardized room acoustics parameters.

The thesis is structured into three main chapters, followed by a final chapter of conclusions. The contents of these chapters will be briefly described here:

- **Chapter 2 - State of the art**
  After some basic definitions of the acoustical theory (in particular, definitions of the acoustical parameters of the ISO 3382 standard), in this chapter, the

outcome derived from the literature search about most considered acoustic environments in the audiological field and already existing databases of scenes are outlined. In addition, in the last section of the chapter, the main characteristics of the Ambisonics audio format are highlighted.

- **Chapter 3 - Ecological audiovisual scenes realization**
  After summarizing the list of environments and scenes suitable for implementing the ecological listening tests, in this chapter the method implemented for recording the AV scenes is illustrated using 2 cases of study. Afterwards, the results from the acoustical characterization of the recorded environments are shown. Then, the second part of the chapter describes the video post-production procedure that has, as its main purpose, the elimination of the tripod that supports the 360° camera.

- **Chapter 4 - Applications: scene reproduction inside the Audio Space Lab**
  In this chapter, the equipment present in ASL is described, and the mathematical workflow to reconstruct the ecological acoustical scenes to perform listening tests are explained.

- **Chapter 5 - Conclusions and future perspective**
  After drawing the main conclusions of the work, this chapter highlights the methodology elements that need to be further optimized in future works, with the purpose to be able to make improvements that will be considered in the future recordings of the next AV scenes to achieve even more realism and immersivity.

Chapter 2

State of the art

To have a basic understanding of the terminology taken for granted in the following paragraphs, some basic definitions of analog and digital acoustics are mentioned in the first section of this chapter. In particular, the definitions of the acoustic parameters described in the ISO 3382 standards [6] [7] are of fundamental importance because they will also be applied in the practical part of the thesis described in the chapter 3.

Subsequently, the obtained results of the two literature searches are described. The first concerns the acoustic environments most considered for the creation of scenes in the audiological field, and the second concerns the databases already present online that have some similarities with the one which will be created for the current project. In a final table, all the environments found in both searches will be summarized.

Finally, the Ambisonics audio format, which will be used for the recordings, is explained. It is a multi-channel and spatialized format, which has the main advantage of being able to record a sound field that can, subsequently, be decoded in different ways to make it adaptable to any array of loudspeakers.

2.1 Fundamentals

2.1.1 Definitions in acoustical field

Before going deep into this thesis's work, some theoretical notions will be defined that will be used hereinafter.

- Sound Pressure Level (SPL): especially used in acoustic physics, it is an indicator of the level of the pressure using as reference value p_{ref}=20 μPa (the hearing threshold at a frequency of 1000 Hz).

$$L_p = 10 \cdot log_{10}\left(\frac{p^2}{p_{ref}^2}\right)$$

Like all level measurements, the unit of measure is decibel (dB, in this case also indicated as dBSPL).

- Decibel Full Scale (dBFS): unity of sound signal level used specially in digital acoustic, which has the reference value A_0 placed at the maximum available amplitude of acoustic signal without peaking.

  $$A(dB_{FS}) = 10 \cdot log_{10}\left(\frac{A^2}{A_0^2}\right)$$

  The maximum value of dB_{FS} is 0.

- Signal-to-noise ratio (SNR): the difference of level between the target signal and the noise (dB).

  $$SNR = L_S - L_N$$

- Speech Recognition Threshold (SRT): SNR at which it is possible for a listener to correctly recognize a certain percentage of the target signal (dB SNR). For example, in SRT80 that percentage is 80%.

- Spatial Release from Masking (SRM): difference between SRT in the co-located condition (both target and masker noise sources located on the same axis of listener's head, so with angles of 0° or 180°) and in a spatially separated condition (with minimum one of the sources in a different direction).

- Room Impulse Response (RIR): the transfer function between an audio signal emitted by a specific position in a room and its version received in another specific position of the same room.

- Digital Audio Workstation (DAW): an integrated software with which it is possible to record, mount, and post-product digital audio; the two main interfaces of it are the mixer to balance the tracks and the timeline to cut and mount them.

- Speech Shaped Noise (SSN): a white noise signal which has the spectrum in frequency with a shape analogous to that one of a long-term speech signal.

- Sweep: a sound signal that spans continuously over a defined frequency range in time (i.e., for a given interval of time only one frequency is present in the signal).

- Binaural Room Impulse Response (BRIR): a RIR calculated simultaneously in two positions that simulate the positioning of the two human ears of a listener. Typically the two microphones used are placed into the ears of a Head and Torso Simulator.

- A-weighted decibel (dBA): decibels weighted in a non-uniform way in frequency. The weight curve used, called A-curve, is consistent with the frequency perception of the human ear and, therefore, gives greater weight to the central frequencies. This unit of measurement is mainly used in studies where human hearing is involved.

- Acoustic scattering: a phenomenon consisting of a not totally specular but partially diffuse reflection of a sound wave on a surface.

- Total harmonic distortion (THD): information that is given as a percentage present in the datasheet of an electronic device related to the distortion introduced by the device itself on the signals that pass through it.

- Dynamic range: is the difference in dB between the maximum level that the electronic component can support without distorting the signal and the level of background noise produced by the component itself.

- Noise floor: in the field of acoustic recordings, is the level of the noise added to the recorded signal and caused by the components of the recording system.

- Head Related Transfer Functions (HRTF): functions that describe how a sound emitted at a certain point is received by human hearing in the two ears. It takes into account several factors, such as shape of the head, ears, and ear canal. These factors also cause a different frequency response from the original sound output.

- Reverberant Tail: decaying sound component due to ambient reverberation still present after the moment when the original sound signal is interrupted.

2.1.2 Room Acoustics Parameters

In order to define the acoustical characteristics of an environment some parameters have been conceived and thoroughly explained in international standards, like the ISO 3382-1 [7] for the measurements of room acoustic parameters in performance spaces and ISO 3382-2 [6] for insights on the reverberation times in ordinary rooms.

All those parameters are computed from in-field measurements of the RIR ($h(t)$) for different spatial configurations of sound source and receiver. In particular, the parameters are computed for each octave band on the decay curve $E(t)$ obtained by applying a backward integration of the square of the $h(t)$.

$$E(t) = \int_t^\infty h^2(\tau)\mathrm{d}\tau$$

Starting from $h(t)$ and the decay curve, the main acoustic parameters of the rooms can be calculated with the methods and formulas described in the standard normative and reported here.

- Reverberation Time: is the time that the decay curve takes to decrease by 60dB after the source emission stops, indicated as $T60$. Since not all environments allow to record a decay of 60 dB, it is possible to derive is from lower range of decays. In particular, the reverberation time can be obtained as: $T = 60/D$, where D is the decay rate of the decay curve obtained by fitting a linear regression line considering a specific decay range. In particular, $T30$ is the reverberation time computed considering the range going from -5 dB to -35dB with respect to the total level of the integrated impulse response; while, $T20$ refers to the range -5 dB to -25.

- Early Decay Time (EDT): estimated SPL decay time of 60 dB computed considering the first 10 dB of decay.

- Early-to-late index: the ratio of the quantity of the early sound energy compared to the late one, considering a specific threshold time t_e (50 ms for speech and 80 ms for music) between the early and the late. C_{50} is also called "Clarity".

$$C_{t_\mathrm{e}} = 10 \cdot \log \frac{\int_0^{t_\mathrm{e}} h^2(t)\mathrm{d}t}{\int_{t_\mathrm{e}}^\infty h^2(t)\mathrm{d}t} \text{ dB}$$

- Early-to-total sound energy ratio: the ratio of the quantity of the early sound energy compared to the total one. An example is the characteristic "definition" D_{50}, also called "Deutlichkeit".

$$D_{50} = \frac{\int_0^{0.05} h^2(t)\mathrm{d}t}{\int_0^\infty h^2(t)\mathrm{d}t}$$

Is easy calculate C_{50} from D_{50} for the mathematical affinity between the two formulas:

$$C_{50} = 10 \cdot \log \left(\frac{D_{50}}{1-D_{50}}\right) \text{ dB}$$

- Center time: time of the center of gravity of the squared impulse response, also defined as the time after that half of the energy of the impulse response is decayed [8].

$$T_s = \frac{\int_0^\infty t{\cdot}h^2(t)\mathrm{d}t}{\int_0^\infty h^2(t)\mathrm{d}t}$$

The T_s is correlated to the reverberation time and increases at its increase [8].

2.2 Literature review of common scenes for hearing research

Research of literature was done to define the scenes to be included in the recordings. This research focused on understanding which environments are most used as a reference in the audiological field and, in particular, taking as the main target the HIOAs.

In [9] seven real-life environments typically frequented by HAs users are identified. One environment, the first, is the reference anechoic laboratory just used for testing hearing aids. The other environments are a street, a supermarket, a cafeteria, a train station, a kitchen, a natural environment, and a panel discussion on a stage. For every scene are reported the noise sources (e.g. other people, cars, environmental sounds), their types (point, located, moving located, or diffuse), and their angular disposition around the listener. There are these data also for the target, but the type is always a speech that can be made from one or more single positions or in a diffuse zone (e.g. the train station). The environments were not real but virtually created with the tool TASCARpro. They were created just acoustically and there are no visual versions of them. Also, other acoustic values like T_{60} or $SPL(dB)$ were calculated in the simulated environment. The RIRs were not calculated in these environments, because the environments were expressly characterized, during their project, with RIRs of similar rooms taken from RIRs libraries. The scenes were made to test algorithms of HA, so the tests were done without any person. The study aimed to demonstrate that testing HA in ecological acoustical environments can be more efficient to predict the efficacy of the HA compared to the standard tests done in laboratory conditions.

Other than this one just seen, in our research we have found a group of articles with some authors in common. Then, the scenes and the environments proposed in different studies have some similitudes but often a different purpose, specific to the single research.

A second paper that shares 2 authors with the one just seen is this one [10]. Here, the experiment aims to obtain patterns of head tracking and eye tracking of normal-hearing young persons in some VR environments, to compare them subsequently with the results of the same test on HIOA or generic HA users. The environments selected were a living room, a lecture hall, a cafeteria, a train station, and a street. This last one had 2 configurations about the role of the listener in the scene (active listener in a conversation or passive listener in a conversation of strangers waiting for someone). In every case, there is the type of elements that make noise with the information also about their movement and location around the listener, and the same information for the targets. Other important data present for every scene are the total $SPL(dbA)$ of the noise and of the target, and the reverberation time T_{60}. To calculate them, have been measured the impulse responses of the simulated environments using a HaTS placed in the position of the listener in the listening system. Every audio scenario is simulated and presents the visual elements with a panoramic video simulated in CG. The setup to do listening tests consists of a system of 28 loudspeakers placed around the listener and controlled with the software TASCAR. The reproduction of the video elements was done with a system of 3 projectors that, together, occupied with their images a field-of-view of 300°.

From the same authors, we must cite also the document [11], because is a preliminary study concerning the other two just presented. Here is explained the possibility of the creation of virtual sound environments with the tool TASCAR, which will be used in the other studies. In particular, in this document, the purposed mentioned is the one developed in the study of the paper [9]: use these simulated sound environments to do tests about the HA efficiency. The environments reported as examples are the cafeteria, the supermarket, the kitchen, nature, and a jazz club. The descriptions of scenarios are less exhaustive compared with the precedent papers, but here there is the jazz club scenario, not present in the other documents. For every scenario, there is a brief description of the elements present, the overall SPL, the degree of diffuseness (DD), and the spatial distribution of the dynamic range. This paper has not a description of any test or hearing setup.

In [12] and [13], two studies made by Zedana, Jürgens, Williges, Kollmeier (author also in [9]) and other authors, the same cafeteria scene was used to test some algorithms of spatial noise reduction for bimodal cochlear implant (CI) users. This type of target uses a CI in one ear and a HA in the other. The scenario is completely simulated and is made using real head-related impulse responses (HRIRs) measured with a HaTS and available online in a database. To prepare for the listening test, the noise and the target signals were convolved with these RIRs. This virtual environment has a defined reverberation time $T_{60} = 1.25s$ and, for the same environment, there are described two different scenarios that have differences in noise type and disposition of noise sources. In every case, the SPL of the noise and the target signal is 65 dB. For every scenario, the positions of the listener, sources of target and noise are described in detail. Were done some SRT tests on bimodal CI users, passing the signal just convolved with the correct RIRIs to their hearing devices. There are no video elements.

In [14], a restaurant virtual environment was developed to test in general people (so the test was done on normal-hearing people) the role of a multi-task condition in the changes in speech recognition in noise and on the efficacy of visual short-term memory. This is one of the less focused on audiology papers among those found (for example, except for the noise SPL, there are no data about acoustical properties of the room like reverberation time or RIRs measured), but the semplicity of the scenario and its adaptability at challenging conditions make it interesting. The environment was created in the same space where the test was done, and the precise dispositions (angles and distances from the listener) of the target and noise sources are reported in the paper. A 2D video was projected on a flat screen and, behind it, five different loudspeakers were placed at the same level where the five virtual target people were projected. The noise source came from a sixth loudspeaker placed behind the listener. The animated video was made with Unity, and the mouth movements were mapped to match their speaking.

Also in [15] has been produced only one environment, but here the study was purely about audiology and HIOA. The purpose was to compare the efficacy of the directionality of HA in three types of environments: real ones, simulated with acoustical models, and simulated listening recordings done in the real ones. To do this was created a cocktail-party scene in a meeting room. This scene was recorded and also re-created with virtual models. On the paper, there is a detailed map of the room with the materials, the characteristics (dimensions and positions) of the target, receiver, and noise sources (7 couples of people talking together). This is the environment more acoustically described among those identified in the research: there are indicated SPL of noise and target, the reverberation time T_{20}, and the critical distance. There are no video elements, so has been possible to rebuild the scene without people but with a human-height placed loudspeaker for every person present in the ideal scene. These loudspeakers emitted anechoic speech. In this environment, were done the HA tests for the case of the real environment, and the recordings for the third type of environment. These recordings were done with the Mixed Order Ambisonics (MOA) technique, recording with a 62-microphone array and, after processing, listening in an anechoic chamber with an array of 41 loudspeakers. In the same chamber were done also the tests for the second environment, the model-based one, simulated with ODEON. To avoid the recordings of all the speech assigned at the target were measured a multi-channel impulse response (mIR) from the target loudspeaker to the microphone array in the listening position. Subsequently, to recreate the target component in the reproduction of the recorded environment, the target anechoic speech signal was convolved with the 62 channel of this mIR.

In a study made in 2014, [16] was realized a new version of a standard audiometric test called "QuickSIN". It is a speech-in-noise test to determine the values of SRT.

The original acoustic material of the test has been modified and listened to in a binaural version with headphones from a group of normal-hearing testers. The scenes used were 8:

1. the original QuickSIN test,
2. the original test with a phase shift of 180° for the target signal in the right ear,
3. an audiovisual version with the vision of the target in a monitor,
4. the audio of scene 2 with the video of scene 3,
5. a spatialized version (target in front direction and noise sources at -90° and +90° at a distance of 1.52m),
6. the scene number 5 located in a virtual room with reverberation time $T_{60} = 0.25s$,
7. scene 6 with an increase of target speech velocity of 50
8. the normal scene with SSN added to the target signal.

To create the spatialized version were utilized the HRTFs of a HaTS. The reproduction of the test was done with a system that used, in addition to the headphones, an audiometer connected to an audio card piloted by a PC with MATLAB.

In 2021, another standard test was revisited in the article [17]. The original test is the Oldenburg Sentence Test (OLSA), which utilizes a matrix sentence test (MST) of words to create random phrases. These phrases are used as target signals to do SRT tests. For this study was created and tested on normal-hearing people an audiovisual version of 150 random phrases assembled from the german female version of the MST. To have coherence with previous studies was maintained the original audio from MST and the videos were created by filming the lip-sync of the same woman who gave the voice at the original audio version of MST. The videos were simply flat 2D recordings in 1080p, with homogeneous illumination and a green screen as background. The listening test was done in a soundproof cabin with an LCD display and binaural headphones.

An underground station virtual environment has been created and described in [18]. This study has a purpose comparable to the one of this project: prototype new standards of hearing tests in some simulated ecological real-life environments. The acquisitions were done in a real station in Munich. The article describes the realization of the environment and the system of reproduction but is not mentioned any test done on it. To make the visual and acoustic models was done a laser scan of the environment, which returned point-cloud data. To view and listen to the acoustic simulated environment, the data acquired were processed with the real-time Simulated Open Field Environment (rtSOFE) software. The visual model was recreated from the union of point-cloud data with some images. For the acoustic model were used two different methods: the image source method to simulate early reflections (using the point-cloud data as a reference to place the planes in the virtual acoustic environment) and multichannel RIRs of the reverberant tail in different positions to get the late reverberations. Were calculated the T_{30} from RIRs measured with an omnidirectional source and some omnidirectional microphones for different octave bands from 125 Hz to 8 kHz. Two different scenes are located in this environment: one with multi-direction talkers around the listener and the second with a sound source approaching or receding from the listener's position. The IRs were recorded in defined positions of source and target with different devices: three types of receivers (omnidirectional microphone, HaTs, and Eigenmike EM 32 microphone array) and two types of emitters (a monitor loudspeakers and an omnidirectional source). Some sound recordings of the real environment were done to reproduce the real noise situation present there. For these registrations, were utilized a close-up microphone for the located noises, and a microphone array for the general noise in some specific moments (e.g. the train passing). In a demonstration, to reproduce everything was utilized Unreal Engine, and the place of installation was a chamber with a system of 61 loudspeakers and the walls covered by projected images.

2.3 Scenes Databases

In this section, the state of the art regarding some existing databases of Ambisonic recordings will be analyzed. Will be shown, for each of these databases, the main differences in comparison to the needs of our project.

2.3.1 ARTE: Ambisonic Recordings of Typical Environments

This database, described in [19], contains the HOA recordings of some ecological acoustic scenes. The purpose for which it was created is completely analogous to the one of this study: to have material that can emulate realistic listening environments in a laboratory so that can be performed hearing tests where the conditions of daily life are truer than those that already exist. These Ambisonic recordings can then be adapted, during playback, to different types of loudspeaker arrays. By doing so, a more realistic reproduction of the environment will be obtained compared to the binaural one used until now in the area of audiometry, because the perceived audio will also vary according to the rotation of the listener's head. Furthermore, Ambisonic recordings of a real environment consider also the interactions that speech has with the surrounding environment, which affects the SNR. This does not happen for traditional laboratory listening tests, which consist in finding a certain SRT in binaural listening of a semi-anechoic speech masked by SSN noise.

The creation of this database was motivated by four basic needs that the authors could not find simultaneously in any other existing database:

1. files in a format that can be reproduced arbitrarily regardless of the loudspeaker array configuration in possession;
2. SPL of the scenes to play them at their original level;
3. acoustic information of the environment and descriptive information about the content of the scene;
4. presence of RIRs to perform speech-in-noise audiometric tests to find levels of SRT.

The environments were selected by choosing from everyday situations characterized by different acoustic conditions, indoors and outdoors, also taking inspiration from the results of previous studies, where were reported the results of surveys concerning the use of hearing aids.

The environments identified are the following:

- library (quiet);
- office (with noise from the workers like typing or talking on the phone);
- church (people talking quietly);
- living room (with loud television sound and noises from the next room);
- church (people talking louder);
- diffuse noise (SSN);
- café (medium occupancy);
- café (internal bar in a company establishment, around lunchtime and out of the room);
- dinner party (babble of 8 people and background ambient music);
- street/balcony (noise from the busy street and inside the apartment);

- train station (at peak hour, with the sound of people, trains, and announcements);
- food court (busy, in university);
- food court (very noisy, in a shopping center at launch time).

Recordings were made with a specially constructed microphone array consisting of a 5cm radius plastic sphere with 62 small capsules inserted. This microphone was placed in the identified environments and recorded life in them for several hours. People who passed through those environments during the recordings were warned about them and every sound source, including people, was kept at a minimum distance of 1.3 m from the microphone. This distance has been maintained in order not to risk, during playback, a distorted sound field caused by the too-close distance. The outdoor recordings were all done in optimal weather, with no wind or rain.

The multichannel RIRs were measured using the same microphone. To calculate them, a Talkbox (the characteristics of this device will be analyzed in the section 3.2) was placed on the horizontal plane of the microphone array, in the frontal position, and at a distance of 1.3 m. This distance was chosen not only for the aforementioned reason of the acoustical distortion but also because it constitutes a realistic distance between two people conversing. The TalkBox emitted, through the reproduction of a MATLAB code, a logarithmic sweep with 20 s of duration, allowing to automatically obtain multi-channel RIR using the signal recorded by the microphone array.

Both the environmental recordings and the RIRs were subsequently processed to extrapolate the HOA tracks. The original tracks are composed of 62 channels, one for each microphone in the array, and the HOA tracks have 31 channels. To switch from the first to the second format, a 31x62 transformation matrix containing the encoding filters was applied to the audio samples. Obtained the HOA files, they can be specifically decoded for playback in any array of loudspeakers. In the case of the study in which the database was presented, the sound environments were reproduced in a spherical array of 41 loudspeakers (this is not relevant to the presentation of the database, but it is helpful to understand how the subsequent conversion to binaural was done).

It is important to observe that the number of HOA channels is apparently anomalous. In fact, it is a complete three-dimensional ambisonics of the fourth order, with 25 channels, to which other 6 channels were added: they constitute exclusively the components located on the horizontal plane of the ambisonics orders ranging from the fifth to the seventh.

Furthermore, were also obtained and loaded into the database the binaural versions of the tracks. To obtain them, a HaTS with two microphones inside the ears was placed in the center of the spherical system of 41 loudspeakers. For both ears, were calculated the HRTFs relative to the positions of all 41 loudspeakers. Then the signals of the 41 channels, each convolved with its two HRTFs (one for each ear, obtaining 82 convolved signals), were grouped according to the target ear and added together to obtain the general channels of the two ears. These binaural files are useful for previewing the listening environments in headphones without the need for a loudspeaker array.

The database has been uploaded online and is publicly available. Each environment has a dedicated directory containing a PDF with the acoustic characteristics of the scene and 2 WAV files: one of the 31-channel HOA version of the recording, and the other of the binaural version. There is also information common to all scenes, such as the MATLAB functions used for various conversions between formats.

This database contains several elements in common with the one that will be created in this study. First of all, the field of use and the objective are the same: it is a database to be used in audiometry to realize hearing tests. Also, the selection of environments is analogous because it considers the results of some previous studies concerning the habits of HIOAs. The file format of the sound environments and RIRs is also more defined compared to the needs of the project of this study, which will use recordings in 3rd order Ambisonics. The main lacks compared to the current project's needs are the complete absence of visual elements of any kind and the presence of only one position of emission for every environment from which the RIR was calculated to the listener position. The current project needs RIRs measured from a lot of emitter points to have the possibility to decide the place of the target or other noises.

2.3.2 Motus dataset

The Motus dataset, described in [20], is a collection of many RIRs related to a single room, with different spatial configurations of the elements present there. It was built because the authors could not find any previously made datasets that simultaneously contained their three main requirements.

1. Physical description of each room configuration, realized using 3D models and 360 photos.
2. Great variety of acoustic characteristics between the different scenarios: there are 830 different configurations, with reverberation times ranging from 0.5 s to 2 s at a sound frequency of 1kHz.
3. Acoustic variety that is caused exclusively by the variation of the quantity, location, and orientation of the furniture: the elements can also possess a complex geometry and can cause specific sound phenomena like scattering.

The main purpose of this database is to study the relationship between the elements present inside a room (type, quantity, disposition, shape, acoustic properties. . .) and the sound field of the room itself. In the specific study done by the authors after the creation of the database, they aimed to correlate the reverberation time of the room with the variation of the quantity of absorption surface present inside. In addition, they have offered to release the database for future research.

The measurements were acquired inside the seminar room of Aalto University, which has dimensions of 4.9 m x 4.4 m x 2.9 m. For each configuration, the positions of the microphone and of the four loudspeakers from which the excitation signal is emitted, have never changed. This disposition is represented in the planimetry present in figure 2.1. The excitation signal consists of a logarithmic sine sweep with a duration of 5 seconds.

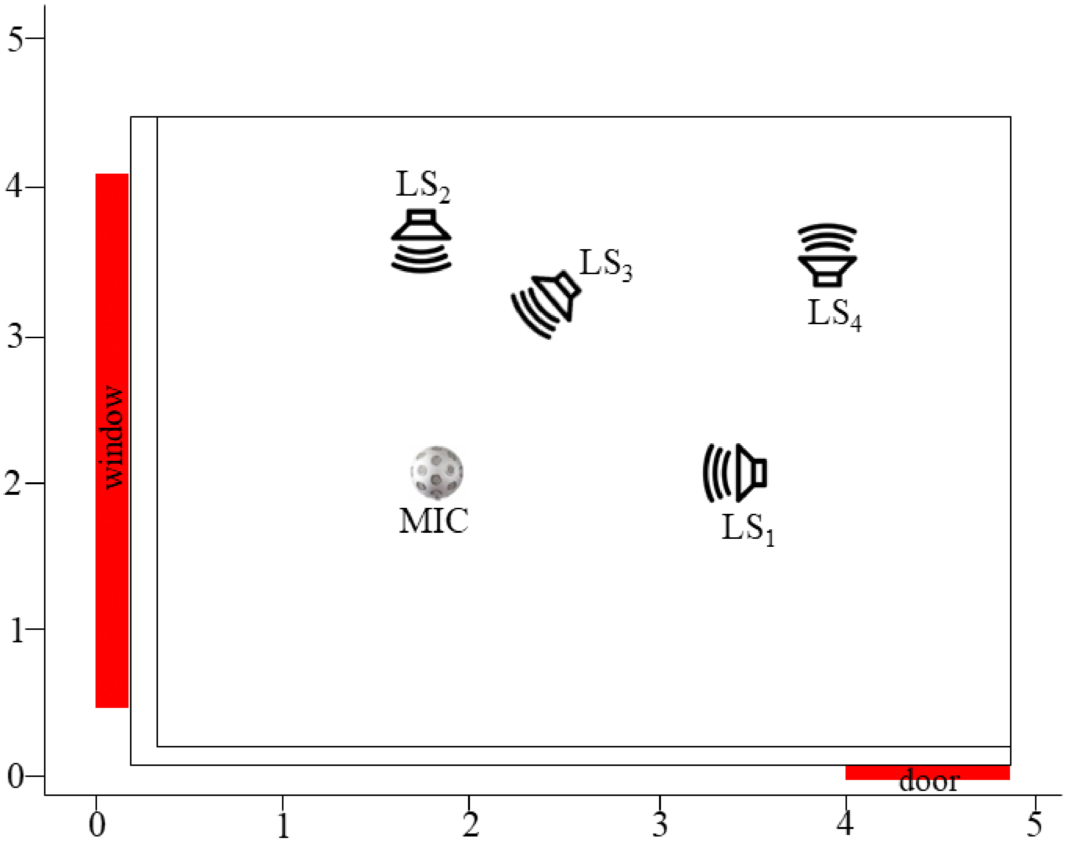


Figure 2.1: Constant positions of loudspeakers and microphone to calculate four different RIRs for every different disposition of the furniture.

There are four types of furniture items available, each with a specific numerical quantity of pieces:

- 6 rollable bookshelves;
- 4 rollable drawers;
- 50 wedges with absorbent surfaces;
- 56 squared carpet tiles with a 0.5 m side.

By changing their number and position, were identified 830 different configurations of the room. For each, were measured the 4 RIRs identified by the positions of the loudspeakers and the microphone, with a total of 3320 different RIRs calculated. For each of the configurations, a photo was taken by a 360 camera placed above the microphone. Finally, the 3D models were created with SketchUp and converted into triangular meshes with Matlab.

The database contains, for each configuration, the following elements in separate files:

- the raw RIRs in 32 channels obtained by the Eigenmike in WAV files at 48 kHz;
- the final RIRs converted to 4th order Ambisonics (25 channels) in WAV files at 48 kHz;
- the 3D model of the entire configuration (room with furniture) as a triangulated mesh available by Matlab or in a .fbx file;
- quantity and type of furniture, with the absorption coefficients of each different surface;
- 3D vectors of the positions and orientations of the microphone and the 4 loudspeakers;
- 360 photos of the configuration in .jpeg and .mat (available by Matlab) format.

Unlike the ARTE database, this one contains some visual elements. In particular, the 360 photos can be considered a previous step compared to the 360 videos that will be done for the current study. Even the fact of making several RIRs, all with the same listening point but with a different position of sound emission, is concordant with the needs of this study. However, this is a dataset of RIRs only, it does not contain ecological complex scenarios that are the basis of the current project.

2.3.3 EigenScape

The Eigenscape database, created for a study done in 2017 [21], is a collection of environmental recordings in Ambisonics created to make research in the field of machine listening, which consists in the computer processing of audio signals to obtain information about the content.

The recordings were made with the mh Acoustic Eigenmike microphone, which consists of a spherical array of 32 microphone capsules able to record up to the 4th order Ambisonics. Recordings were made at 24-bit sample resolution and with a sample rate of 48 kHz, keeping the order of tracks in accordance with the Ambisonics Channel Number standard. All recordings were made with the EigenStudio acquisition software with the gain set at +25dB, with the exception of the track 'TrainStation-08', on which was set a gain of 5dB because it contained high-level sounds that would clip at a higher gain. All recordings, except one made indoors, were made with a windshield on the microphone. To make the recordings, the microphone was mounted on a stand (tripod or monopod depending on the environment of the registration) and also a Samsung Gear 360 video camera was mounted on the same stand. The video content will be used exclusively to identify the events that happened during the scenes when these are incomprehensible from audio information alone.

Eight different scenes, in this study defined as "classes", have been identified, and for each of them were made eight different recordings. The total number of recordings is therefore 64. Each one has a duration of 10 minutes. The main requirement regarding which the classes were decided was the ease of recording in the environment and the variety of acoustic environments between the different scenes. The eight classes identified are the following:

- beach;
- busy street;
- park;
- pedestrian zone;
- quiet street;
- shopping center;
- train station;
- woodland.

A few seconds were cut at the beginning and at the end of the recording tracks to clean the registrations from non-ambient noises produced by the operators to start and stop the recordings. In general, non-diegetic sounds correlated to the act of record, such as some curious people asking questions about what was going on, were minimized as much as possible. In the most crowded environments, these accidental interventions weren't a big problem because they were covert in the general speak-shaped noise of the environment.

The database is available online for download and is structured in ZIP files, each of which contains the registrations of the specific class in uncompressed WAV files. Each ZIP, therefore, contains 8 files of 10 minutes each with 25 channels, and the total size of the database is 140 GB. The database metadata contains information regarding the physical description of the environments and the type of stand used (monopod or tripod depending on the needs of the individual environment). Other elements contained in metadata are the dates and times of the recordings and the presence of the windshield. Due to the excessive data size, is available a second and more manageable version of the database with a total size of 12.6 GB. This contains all the tracks of the first version, but just in first-order Ambisonics and with a FLAC type lossless compression.

Like the other databases, also this reflects only a few of the needs of the current project. It is a collection of Ambisonic recordings of real environments that set up complex sound environments. There is also a video component, which however is used only as a reference to verify what is happening in the environment. The other needs are not considered here, especially because the database was not made for research in the field of audiometry, but for machine listening. No RIRs were detected in the environment and no attention was dedicated to the environmental acoustic parameters. It can be used as a reference for practical advice on how to make recordings, especially regarding the advice to trim the beginning and end of clips, to minimize non-diegetic noise, and to be prevented against the risk of interfering with strangers who happen to be in the recording site.

2.3.4 Auditory-visual scenes for hearing research

The scene of the underground station described before in the article [18] is part of a dataset realized by the same authors and described in [22]. This dataset was created to have ecological scenes for hearing tests that reflect the listening conditions of everyday life, so it has a purpose analogous to the one of the current study.

Originally three starting environments were created: an underground station, a pub, and a living room. For each environment there are two scenes, meaning as a scene a specific disposition of sound sources and receivers. In paragraph 2.2 there is a description of elements acquired in the physical place to recreate the environment acoustically and visually. It will be therefore avoided going into detail here on the aspects of acquiring in the environments, already described previously.

The database offers the possibility, for external researchers, to add environments created by them, but they have to respect the standard conditions established by the creators. There is a defined document structure called the "Environment description document" in which all the useful characterization data of the environment must be inserted. The first part of this document contains the description of the environment, with information relating to the location (place, physical dimensions, acoustic description of the environment), the floor plan (including the positions and orientations of sources and receivers), recordings made of the sound scene, acoustic measurements made, acoustic and visual models made. The second part contains a description of the structure of the technical files present in the environment's folders, explaining the roles of the files and how to use them to correctly reproduce the environment.

It is required to use the .obj format for 3D objects, with textures in separate files referenced in the .obj file or associated with materials contained in .mtl files. Another recommended format is the SOFA (spatially oriented format for acoustics), which is used to describe the directivity of the sources and the receiver. Finally, it is recommended to add some HRTFs related to points in the space where measurements have been performed with a HaTS. They will be useful to understand if the BRIRs simulated in the virtual environment give results coherent with the HRTFs recorded in the real environment.

Can be learned an example from this database on the audio side. The level of detail of the measurements done in the rooms allows for a very faithful and versatile acoustic reconstruction of the sound environment like the one wanted to achieve in the current project. From a visual point of view, the static model has the limitation of not making visible the dynamism of the places. For this reason, unlike what was done for this dataset, will be made video recordings in this study.

2.4 Speech intelligibility: influencing factors in complex environments

One of the objectives of this thesis will be to find some challenging listening conditions for older adults. To identify them they must be considered the results of studies concerning the intelligibility of speech in various noise conditions, especially in environments used for functions where speech has a significant weight. Speech intelligibility depends on various factors related to the noise source (distance, direction from the target, type of noise) and the environment (such as the presence or absence of acoustic treatment).These results will be considered to define the positions of the target signal and the noise in the identified scenes.

The author Bronkhorst, in the article [23], made a review of his research in the literature concerning the cocktail party phenomenon. The phenomenon consists of the negative influences on speech intelligibility caused by the presence of one or more speakers different from the target. Concerning the scene task, other speakers are sources of noise that increase the perceived complexity of the scene and generally decrease the intelligibility of the target speech.

Within this review, various elements that affect speech intelligibility are defined and have been reported in the following list

- Spectral differences: the similarity in the time-averaged frequency spectrum between the target signal and the masking noise positively compromises the effectiveness of the masking. In particular, the change in SRT can vary between -2dB and +2dB depending on this factor.

- Fluctuations: a large amplitude of the changes in the sound pressure level of the signal over time contributes to an increase of the release from masking, with a decrease of the SRT from 6db to 10dB compared to what happens with a stationary noise. These advantages decrease when the number of different noise sources increases.

- Voice similarity: in the case of a masking signal consisting of a single speaker, the similarity of this voice with that of the target signal helps to raise the SRT from 3 dB to 9 dB. The most disadvantageous case is the one in which the two voices belong to the same speaker.

- Spatial separation: as already mentioned in the paragraph 2.1.1, spatial release from masking (SRM) is a phenomenon that consists of the greater intelligibility of a target signal when the noise source is spatially separated from it. This phenomenon decreases as the number of noise sources increases, with a maximum decrease in SRT of 11 dB in the case of a single noise source and 8 dB in the case of multiple sources.

- Best ear or binaural listening: with the same stimuli, binaural listening helps to significantly increase SRM thanks to the phenomena of head shadow and interaural time difference. The decrease in SRT brought about by this condition ranges from 1 dB to 7 dB.

- Reverberation: the acoustic condition of the environment is closely linked to speech intelligibility, with an increase in SRT of up to 9 dB for longer reverberation times.

- Divided attention: studies on this aspect have not identified a precise value of the variation of the SRTs, but it has been found that, usually, tasks that require the use of divided attention (concentrated on multiple elements simultaneously) lead to worse performance (and, consequently, higher SRT) than those requiring the use of selective attention (concentrated on a single specific element).

- Moderate hearing impairment: in a contest such as that of the cocktail party, the benefit brought by the HAs is limited because they amplify the target and the noise sources in the same way. All the advantages seen in the previous points are therefore attenuated, causing an SRT value up to 10 dB higher than the same environmental conditions experienced by a normal-hearing individual.

In a 2019 study made by Puglisi et al. [24] speech intelligibility was measured in two classrooms in the city of Turin. The goals of their experiment were two:

- to measure and report the changes of the SRT80 under different acoustic treatment conditions and with different types of noise;
- to report the changes in SRM depending on the reverberation of the room and the type of noise.

To satisfy the first objective, the measurements were made in two different classrooms: the first has been previously treated to make it compliant with the acoustic standards, and the second was not treated. Consequently, there will be a different reverberation time and a different level of speech intelligibility. Were also used different types of noise. The classification of the type of noise is made by defining the differences between energetic noise and informational noise. Unlike energetic noise, informational noise has semantic contents that interfere with the semantic information of the target, further limiting its intelligibility. In this case, a stationary SSN was used as energetic noise and a children's babble recorded under semi-anechoic conditions was used as informational noise.

To achieve the second objective, however, the measurements were made by placing the noise source at different distances and angles for each position of the receiver.

On-site measurements of the monaural RIRs of the chamber were performed to obtain the reverberation time $T30_{0.25-2kHz}$ and clarity $C50_{0.5-1kHz}$, but also of the BRIRs with the aim of convolve them, in the listening phase, with their respective target and noise signals.

To measure the RIRs, an omnidirectional microphone was used in the receiver positions and a TalkBox (the characteristics of this device will be analyzed in the section 3.2) in the point of emission of the target signal. It is important to clarify that, in this study, the receiver and target positions were realistically decided.

The microphone was placed at a height of 1.1m from the floor in positions of the classroom where the student desks are located, and the Talkbox was located in the position occupied by the teacher at the desk at a height of 1.5m from the floor. Figure 2.2 shows, in the floor plans of the classrooms, the locations used for the microphone (receiver, R), and for the Talkbox (target, T) during the measurements of RIRs.

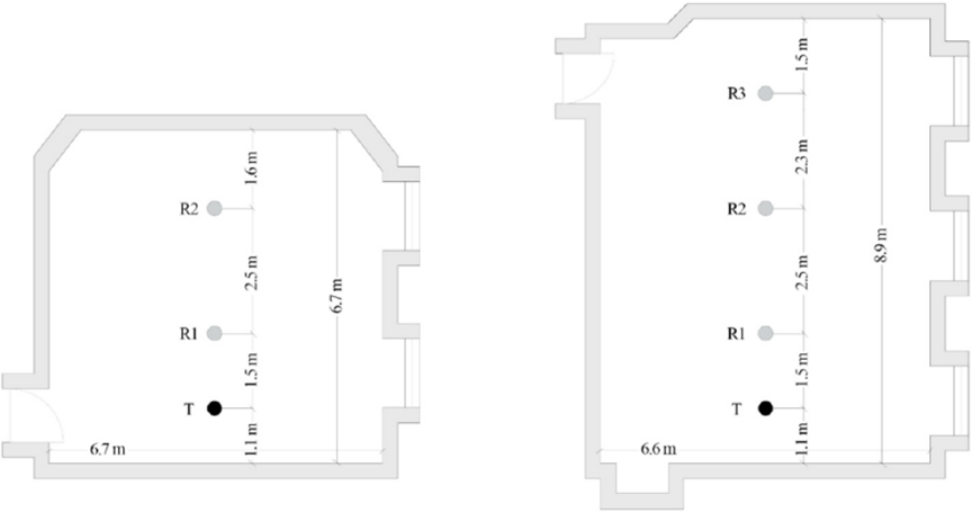


Figure 2.2: Measurements of RIRs done, with locations of target and receivers in the compliant classroom (left) and in the non-compliant (right) [24].

To measure the BRIRs, however, partially different equipment was used. For the target speech sources, the same TalkBox of the previous measurement was used, but at the receiver position was placed a HaTS with two microphones located inside the ears. Instead, an omnidirectional dodecahedron-shaped emitter with 12 loudspeakers was placed at the points of the noise sources.

Twenty-three BRIRs were measured: 5 for each receiver position and 18 for each noise position (M for masking in figure 2.3).

Subsequently, was done the listening test phase. Noise and target signals were convolved with their respective BRIRs. The sentences of the SiIMax (simplified Italian Matrix) were used as the target signal. The SiIMax is a simplified version of a matrix, ITAMatrix, used to randomly generate sentences in Italian with syntactic coherence but no semantic meaning. In the case of the simplified version, the one used to generate the target signals, only number-name-adjective triads are generated. The target and noise signals were convoluted with the respective BRIRs and then listened to by 43 volunteers in binaural headphones. For each listening test, the goal was to define the SRT80. To do this, the noise was always kept

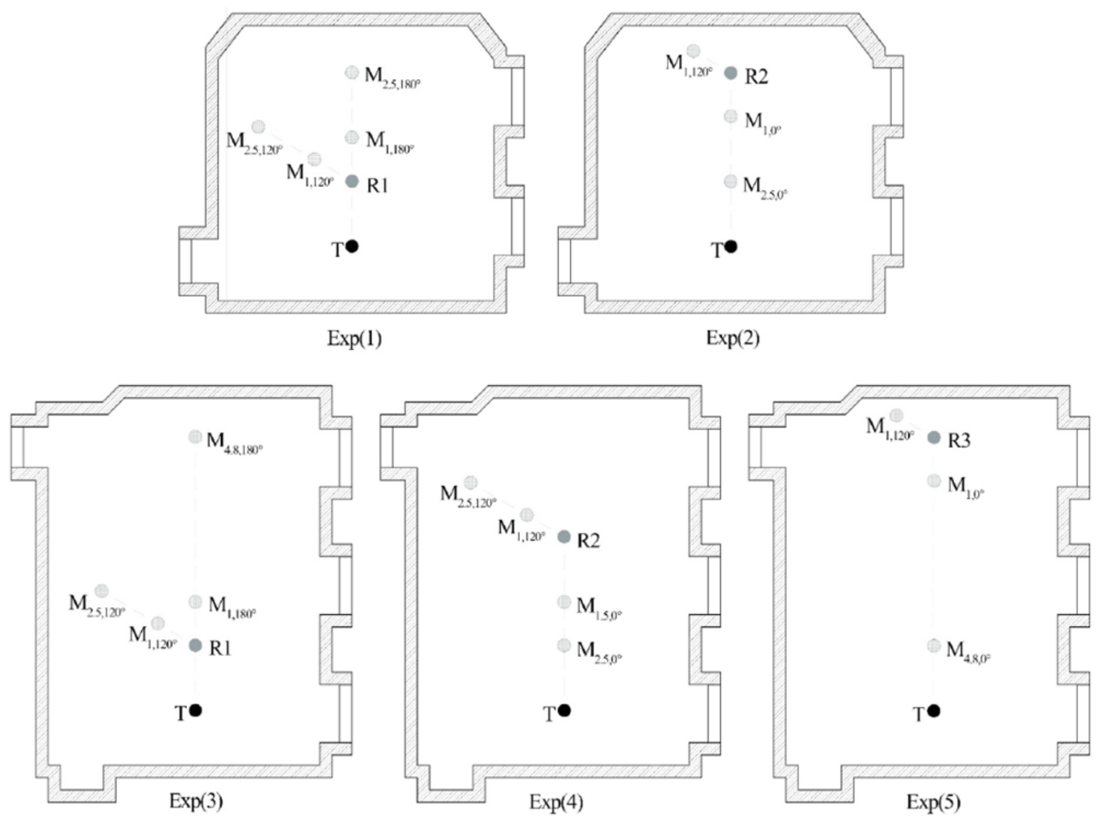


Figure 2.3: Places of masking noise sources where BRIRs were measured for every target-listener couple identified in the previous figure. [24]

constant at 60 dB and the target level was varied until was found the right one in which the number of words recognized by the listener was equal to 80% of the total.

After the various analyses made on the results, the conclusions were summarized in the following points, which report the average speech intelligibility behavior in relation to the considered variables.

1. A longer reverberation time negatively affects speech intelligibility. In the non-compliant classroom, on average, was reported an SRT80 higher of 6 dB for informational noise and 5.4 dB for energetic noise.

2. Speech intelligibility decreases at the increase of target-receiver distance. Passing from 1.5 m to 4 m of distance, an average increase of 3dB of the SRT80 is signaled, from which it is possible to obtain an estimate of an increase of about 2 dB when the distance is doubled.

3. The type of noise affects speech intelligibility, with the result of a decrease of

it in the presence of informational content in the noise. Considering all the measurements made, on average the SRT80 is higher by 7 dB in presence of informative noise than in presence of energetic noise.

4. SRM is relevant in two cases: with energetic noise in the low reverberation room and with informational noise in the high reverberation room. In both cases just mentioned, the SRM value is about 3 dB in the case where the noise-receiver distance is 1 m. In the case of a noise-receiver distance of 2.5 m, the SRM becomes negligible with values that oscillate between 1 dB and -1 dB.

The results of this study will be useful in the phase of the decision of the environments, because they correlate the intelligibility of speech in an environment to the various parameters just seen and, consequently, the perceived acoustic complexity (also called "challenging") of the environment itself. Furthermore, the equipment used by the authors of this study is available at the Polytechnic of Turin, and it will therefore be possible to use it to calculate the RIRs in the environments chosen for the measurements in the current project. This can be done with speech and noise emitters, but not with the acquisition microphones because the RIRs in our environments must not be binaural but Ambisonic. Finally, in the selection of the environments, the two classrooms described will be taken into consideration, because thanks to this previous study their main acoustic features are already known.

2.5 Summary of the environments found

The following table shows the environments found in the research done about scenes and databases (the results of them have been shown in the paragraphs 2.2 and 2.3). Environments are ordered from most addressed in research to least, citing all related papers.

2.6 Ambisonics soundfield encoding technique

2.6.1 Main concepts & Spherical harmonics

In this section, we will explain the basic principles of Ambisonics, i.e. the audio format that will be used here for the recordings.

Ambisonics is a spatialized audio format, so it considers the directivity of recorded sounds. In particular, Ambisonics can be considered as a technique of recording the entire sound field [25] around a central point where the microphone is placed. As it will be explained in more detail, the recordings of the various

Environment	Articles
Cafeteria	[9] [10] [11] [12] [13] [19]
Train station	[9] [10] [19] [21]
Street	[9] [10] [19] [21]
Supermarket	[9] [11] [21]
Nature	[9] [11] [21]
Living room	[10] [19] [22]
Anechoic laboratory	[9] [19]
Restaurant / pub	[14] [22]
Kitchen	[9] [11]
Panel discussion / lecture hall	[9] [10]
Jazz club	[11]
Cocktail party	[15]
Audiovisual QuickSIN test	[16]
Classroom	[24]
Library	[19]
Office	[19]
Church	[19]
Dinner party	[19]
Food court	[19]
Beach	[21]
Park	[21]
Pedestrian zone	[21]
Underground station	[22]

Table 2.1: The environments found are sorted in descending order with respect to the number of citations.

microphone capsules are subsequently encoded in the Ambisonics B-format, which can be then decoded differently based on the loudspeaker array that is used for the final listening. Since it is a detection of the sound field in all directions, Ambisonics is mainly used when there is the need to reproduce the entire sound field around the listening user, as in a virtual reality installation (e.g. the Audio Space Lab at Polytechnic of Turin).

To understand better how sound field coding works in Ambisonics it is necessary to introduce the concept of spherical harmonics. To easily understand this concept, it will be compared with its two-dimensional equivalent.

It's assumed that a generic waveform $s(t)$ is sampled using the PCM (pulse-code modulation) technique, therefore sampling it at regular intervals of time and quantizing the amplitude in values depending on the bit depth established for the digitization of this signal. The sampled signal can subsequently be analyzed with Fourier analysis: this analysis makes the signal visible in the frequency domain and represents it as the sum of some sinusoids of different frequencies, each one with a specific amplitude and phase [26].

Spherical harmonics are the three-dimensional equivalent of these sinusoids just seen. A sound field can be considered, in a single instant, modulated in PCM: only some equally spaced directions will be considered (similarly to the time discretization of the monodimensional signal) and, to each of these directions, a quantized sound pressure is attributed, with precision depending on the bit depth. This signal can also be represented as the sum of a set of spherical harmonics, each one with a specific weight (similar to the amplitude assigned to sinusoids) [26].

A visual example to better understand the concept is given in figures 2.4 and 2.5.

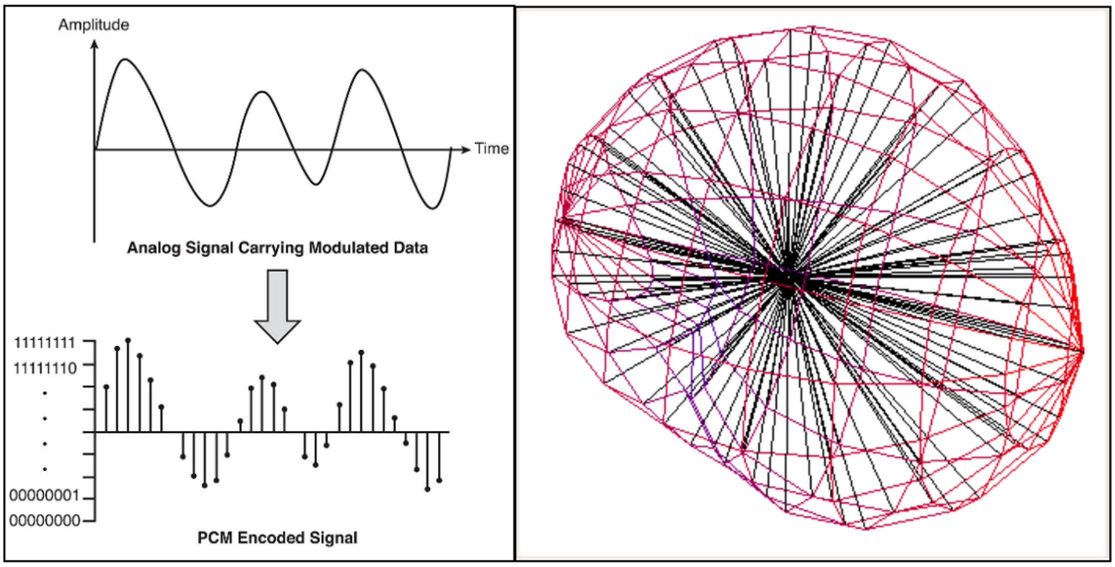


Figure 2.4: Temporal and spatial PCM [26].

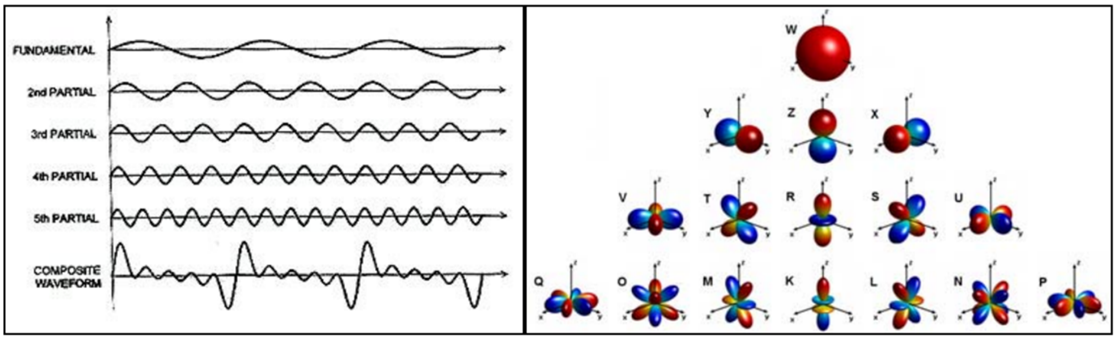


Figure 2.5: Analogies between Fourier analysis in time and spherical-harmonics analysis in space [26].

2.6.2 Ambisonics orders & formats

Depending on the desired level of detail of the recreated sound field, there are different orders of Ambisonics, each one with a different number of signals equivalent to the number of spherical harmonics used.

In general, the following rule relates the order number n to the number of channels used N [25]:

$$N = (n + 1)^2$$

The 0-order Ambisonics contains only one omnidirectional channel, and the first order that discretizes the different directions of different sounds is the 1st, which is also the most used order. Orders superior to the first are called "Higher Order Ambisonics" (HOA) and contain greater detail and directional precision by increasing order. The one that will be used for this project is the 3rd order Ambisonics, with 16 channels.

An Ambisonics track can be represented in several multi-track formats, identified by letters of the alphabet. The two main ones that must be considered for the following explanations in this chapter are formats A and B.

- A-format: this is the output signal of Ambisonics microphones, with each track corresponding to a microphone capsule. In the simplest case, the one of the 1st order, the four tracks are the sound pressures recorded by four microphone capsules disposed on the sides of a tetrahedron (Soundfield microphone). This format varies according to the microphone, because the capsules, ideally placed in the same position but oriented in different directions, in reality, have a variable distance from the center of the system depending on the microphone model [27].

- B-format: it is the encoded version of the A-format and recognized as the universal one; it is independent of the microphone used and can be decoded in different ways according to the disposition of the loudspeakers in the listening array. The channels of this format are dedicated to the weights of the components of the sound signal in the directions described by the single spherical harmonics, therefore each harmonic has its dedicated channel. Also for this format, it is easier to concretely understand the contents of the different tracks by explaining the ones of the 1st order. They are 4: the first, called W, contains the omnidirectional information already present in order 0 and the other three, called X, Y, and Z, contain information relating to the pressure gradient along the three Cartesian spatial directions[27].

2.6.3 First-order Ambisonics: coding and decoding

As we have already mentioned before, the 1st order Ambisonics is the one with which it is possible to explain the encoding and decoding process in a more concrete way. Coding means the transition from the A-format output of the microphone array to the representation of the sound field in the B-format. By decoding, we mean the transition from the B-format to the set of tracks dedicated to the single channels of the listening loudspeakers array.

To explain the encoding is efficacious to start with an example of a general sound field with the sound components $s_i(t)$ and, theoretically, the four B-format channels can be obtained with the following conversions [25]:

$$\begin{aligned}
W(t) &= \sum_i s_i(t)/\sqrt{2} \\
X(t) &= \sum_i s_i(t)\cos\phi_i\cos\delta_i, \\
Y(t) &= \sum_i s_i(t)\sin\phi_i\cos\delta_i, \\
Z(t) &= \sum_i s_i(t)\sin\delta_i.
\end{aligned}$$

The W channel contains the sum of all the components $s_i(t)$ equally weighted and normalized by the constant factor $1/\sqrt{2}$, and the X, Y, and Z channels contain the sums of the projections on the respective axes of the sound sources pressures, depending on the direction in which they are oriented (ϕ_i and δ_i are, respectively, the azimuth and the elevation angles of a single sound source). The equalization done in the W channel is according to a standard, called "Furse-Malham" (FuMa), of normalization coefficients given to every component of a specific order of Ambisonics [27].

These theoretical formulas can be applied to a microphone array with several capsules, setting the various $s_i(t)$ equal to the outputs of the single microphone capsules. However, it is necessary to apply some corrections to the X, Y, Z, and W outputs because the capsules are not coincident, neither with each other nor with the center of the system. The theoric center of the system is the listening point to which the formulas just described refer.

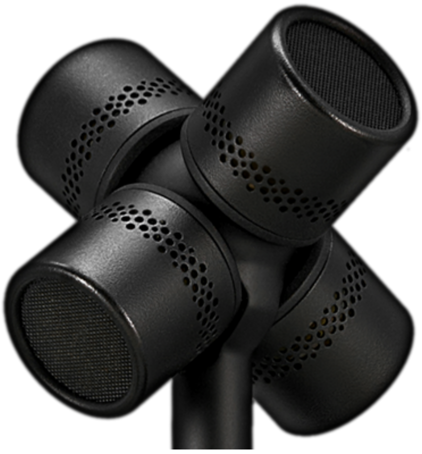


Figure 2.6: An example of Soundfield microphone.

A typical example is that of the Soundfield microphone (Fig 2.6), already mentioned above because it is capable of recording in Ambisonics A-format. It is a tetrahedral microphone array. Its four capsules can be named in the following method that reflects their spatial disposition: left-front–up (LFU), right-front-down (RFD), left-back-down (LBD), and right-back-up (RBU). As just mentioned, the formulas to encode from the A-format made by this microphone to the B-format will have to receive frequency-dependent compensations because the frequency response of this system is distorted, especially for frequencies where the wavelength is less than the distance of each capsule from the center of the system.

The formulas for coding, in this case, are as follows [25]:

$$
\begin{aligned}
W &= F_0(\omega)(LFU + LBD + RFD + RBU)\\
X &= F_1(\omega)(LFU - LBD + RFD - RBU)\\
Y &= F_1(\omega)(LFU - RBU - RFD + LBD)\\
Z &= F_1(\omega)(LFU - LBD + RBU - RFD)
\end{aligned}
$$

$F_0(\omega)$ and $F_1(\omega)$ are two compensation formulas in frequency and phase, respectively for the omnidirectional channel and for the three sound pressure gradients along the axes. They vary depending on the microphone used for the recording.

In the decoding phase, the signal to assign at the single loudspeaker is calculated by solving the following linear system:

$$\begin{pmatrix} W(t)/\sqrt{2} \\ X(t) \\ Y(t) \\ Z(t) \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \cos\phi_1\cos\delta_1 & \cos\phi_2\cos\delta_2 & \cdots & \cos\phi_N\cos\delta_N \\ \sin\phi_1\cos\delta_1 & \sin\phi_2\cos\delta_2 & \cdots & \sin\phi_N\cos\delta_N \\ \sin\delta_1 & \sin\delta_2 & \cdots & \sin\delta_N \end{pmatrix} \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{pmatrix}$$

Here, ϕ_n and δ_n are the angular positions of a single loudspeaker, and $s_n(t)$ is the signal assigned as input to that loudspeaker. Every channel of the B-format must be multiplied by his FuMa coefficient.

2.6.4 Introduction to Higher Order Ambisonics

In this case, will be utilized the Ambisonics of the 3rd order. In figure 2.7 there are the physical representations of the directivity patterns of every component of the first six orders of Ambisonics.

Order	Channel	SN3D	FuMa Weight
0	W	1	$1/\sqrt{2}$
1	X	$\cos\phi\ \cos\delta$	1
	Y	$\sin\phi\ \cos\delta$	1
	Z	$\sin\delta$	1
2	R	$(3\sin^2\delta$ - $1)/2$	$2/\sqrt{3}$
	S	$(\sqrt{3}/2)\ \cos\phi\ \sin(2\delta)$	$2/\sqrt{3}$
	T	$(\sqrt{3}/2)\ \sin\phi\ \sin(2\delta)$	$2/\sqrt{3}$
	U	$(\sqrt{3}/2)\ \cos(2\phi)\ \cos^2\delta$	$2/\sqrt{3}$
	V	$(\sqrt{3}/2)\ \sin(2\phi)\ \cos^2\delta$	$2/\sqrt{3}$
3	K	$\sin\delta(5\sin^2\delta\text{-}3)/2$	1
	L	$(\sqrt{3}/8)\ \cos\phi\ \cos\delta\ (5\sin^2\delta\text{-}1)$	$\sqrt{45}/32$
	M	$(\sqrt{3}/8)\ \sin\phi\ \cos\delta\ (5\sin^2\delta\text{-}1)$	$\sqrt{45}/32$
	N	$(\sqrt{15}/2)\ \cos(2\phi)\ \sin\delta\ \cos^2\delta$	$3/\sqrt{5}$
	O	$(\sqrt{15}/2)\ \sin(2\phi)\ \sin\delta\ \cos^2\delta$	$3/\sqrt{5}$
	P	$(\sqrt{5}/8)\ \cos(3\phi)\ \cos^3\delta$	$\sqrt{8}/5$
	Q	$(\sqrt{5}/8)\ \sin(3\phi)\ \cos^3\delta$	$\sqrt{8}/5$

Table 2.2: Expression to mathematically define the spherical harmonics of Ambisonics until the 3rd order.

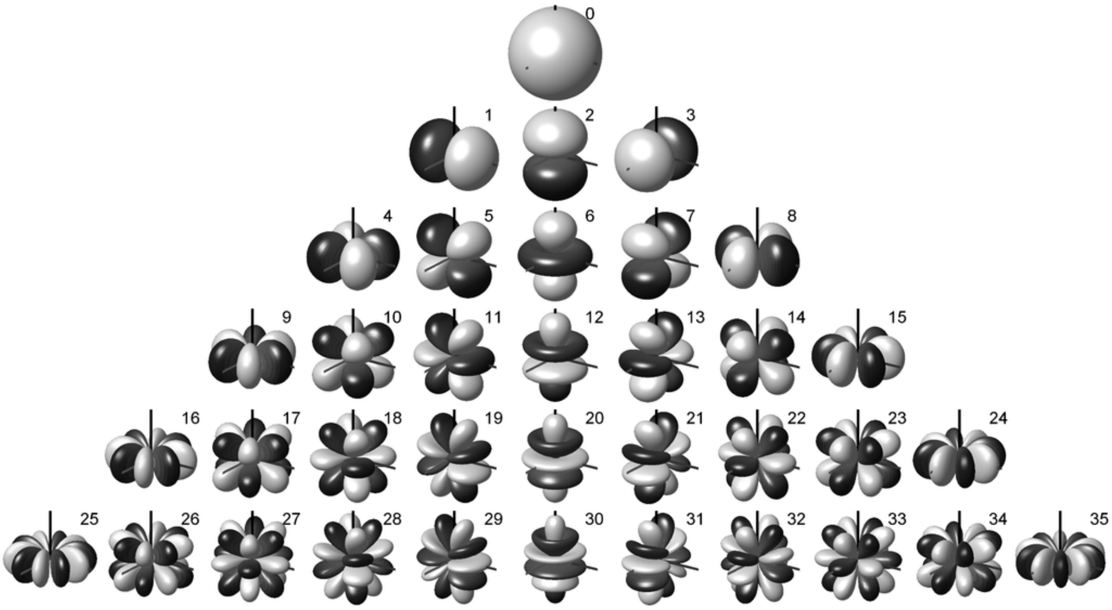


Figure 2.7: A representation of spatial patterns of the spherical harmonics of Ambisonics orders until the 5th. The light zones are the positive components, and the dark ones are negative. [28]

In the table 2.2 are reported the expression to describe the spherical harmonics of HOA until the 3rd order as they are represented in [27]. These expressions are called SN3D because they consist of the expressions of the spherical harmonic normalized in a special way called "Schmidt semi-normalization". The process to encode and decode HOA is analogous to the one just described for the 1st order, but with more channels to encode and more equations to solve in the system of the decoding phase.

Chapter 3

Ecological audiovisual scenes realization

In light of the environments found in the previous chapter 2.1, in this chapter some of them are identified to be included in the current project. Then, two of them are selected to start acquiring the AV scenes, i.e., a conference hall and a classroom. Thus, the recording procedures used to capture scenes inside those environment are then described. In particular, after a descriptive overview of the main devices used during the measurements, the workflow performed in the environments is described.

The data obtained from the measurements are the acoustic parameters of the environments, the multichannel RIRs which will be used to reconstruct the sound scenes, and the 360° video footage of all the individual scenes set in those environments.

The second part of the chapter describes the video post-production procedure that allows the elimination of the tripod that holds the camera in 360 videos.

Before starting to discuss the defined environments and the scenes created in them, it is necessary to make a recap of the characteristics of the dataset that will be created in the current project. This dataset will be created to be able to simultaneously respect all the requirements of the current study, requirements that have not been found satisfied simultaneously in any of the datasets already created.

- Acoustic and visual realism of the created scenes;
- visual content seeable on a 360 viewer, but compatible with all main types of video playback devices;
- environments often frequented by HIOA, in which they have a perception of acoustic complexity;

- a database of RIRs recorded in every single environment to decide, in the listening phase, the positions of target and noise sources with a good possibility of choice;

- acoustic characterization of the environment, with the main parameters of the standard ISO 3382 known;

- audio tracks in a loudspeakers-independent format (preferably HOA).

3.1 Scenes selection

The first confirmed environments in which the scene measurements will be performed are those present in the table 3.1. The selection is based on what has been found in the literature, and the scenes built in these environments will be created in a way based on the results of the studies analyzed in the paragraph 2.4, so considering the effects of the types of noise and the dispositions of the noise sources on speech intelligibility. It is important to specify that the acoustic data reported in this table are not the real ones of the environments, but are estimated based on what is reported in the literature regarding similar environments. The real ones will be calculated through measurements of IRs in the specific location.

Two other specifications need to be mentioned to better understand the table: the type of noise and target is mentioned just when are present in the original article, and the level refers to SPL in the considered position of the receiver (*); some environments are characterized by a mixed ambient noise with informational and energetic components (**).

In the next paragraphs, will be described the measurements performed in the first two environments in which they were made. These environments are a conference hall and a classroom, decided firstly because the locations used for them in the current study are rooms in which is known that there is a high reverberation time. This causes more perception of acoustic complexity in the environment. In addition, the room amplification system is available, so that it is possible to compare measurements performed with the target speaking directly or amplified.

Environment	Reverberation condition	Noise level and type*	Target level and type*
Laboratory (reference A) [9]	Anechoic	Located SSN 65dB	Located speech 65 dB
Laboratory (reference B) [29]	Anechoic	Diffused SSN 58.3 dB 54.2 dBA	Located speech 67.2 dBA
Cafeteria [29]	$T_{30} = 1.1$ s	** 71.0 dB 67.3 dBA	Located speech 72.1 dBA
Food court [29]	$T_{30} = 0.2$ s	** 78.2 dB 74.9 dBA	Located speech 74.9 dB
Office [29]	$T_{30} = 1.1$ s	** 56.7 dB 51.4 dBA	Located speech 63.9 dB
Living Room [29]	$T_{30} = 0.2$ s	** 63.3 dB 58.7 dBA	Located speech 65.0 dB
Conference Hall [10]	$T_{60} = 0.78$ s	Diffused energetic 44.6 dBA	Located speech amplified by two loudspeakers 51.7 dBA
Classroom (non-compliant) [24]	$T_{30} = 3.1$ s	Variable in type and level to do SRT80 test.	Located speech 60.0 dB

Table 3.1: The selected environments in which to make the measurements.

3.1.1 Conference Hall

The conference hall where the measurements were made is located in the Egyptian Museum of Turin.

The place from where the speaker makes the speech was used as the place of the source S. The two receiving points R_1 and R_2 were placed on the chairs in line with the source and at a distance from it of 4.1 m and 9.8 m respectively. The noise sources, that always maintain a distance of 1.8 m from the listener, are directive one-talker type. Noise sources that do not lie on the axis between the source and the receiver are displaced at an angle of 120° from it. It is important to specify that the first number subscript of the noise source N refers to the number of the receiving position R from which the noise is heard. The second, however, is a numbering of the noise sources that refer to the same R. The angles and distances of the loudspeakers from the receivers are shown in the table 3.2.

Receiver	Angle LS1	Distance LS1	Angle LS2	Distance LS2
R_1	-65°	4 m	66°	4.2 m
R_2	-26°	8.2 m	27°	8.3 m

Table 3.2: The relative arrangement of the loudspeakers to the two reception points in the conference hall.

In the video content, the noise source is represented by a HaTS, while some extra people are sitting scattered around the room, impersonating other spectators of the conference. There are 14 recorded scenes in total. To the 5 scenes with noise defined in the figure 3.1 are added the two in the points of the receivers with only the target without noise. These scenes are, in turn, multiplied by two because they were measured both with and without a microphone.

3.1.2 Classroom

The classroom where the measurements were made is the 1T at the Politecnico di Torino.

Similarly to what how seen for the previous environment S, R_1, and R_2 lie on the same axis.The noise-receiver distance is constant at 1.9 m, but the type of noise is a multi-talker SSN. Here too the angles of the misaligned noise sources are 120° and the identified scenes are, in all, 12 (to be counted in the same way seen for the conference hall). The angles and distances of the loudspeakers from the receivers are shown in the table 3.3.

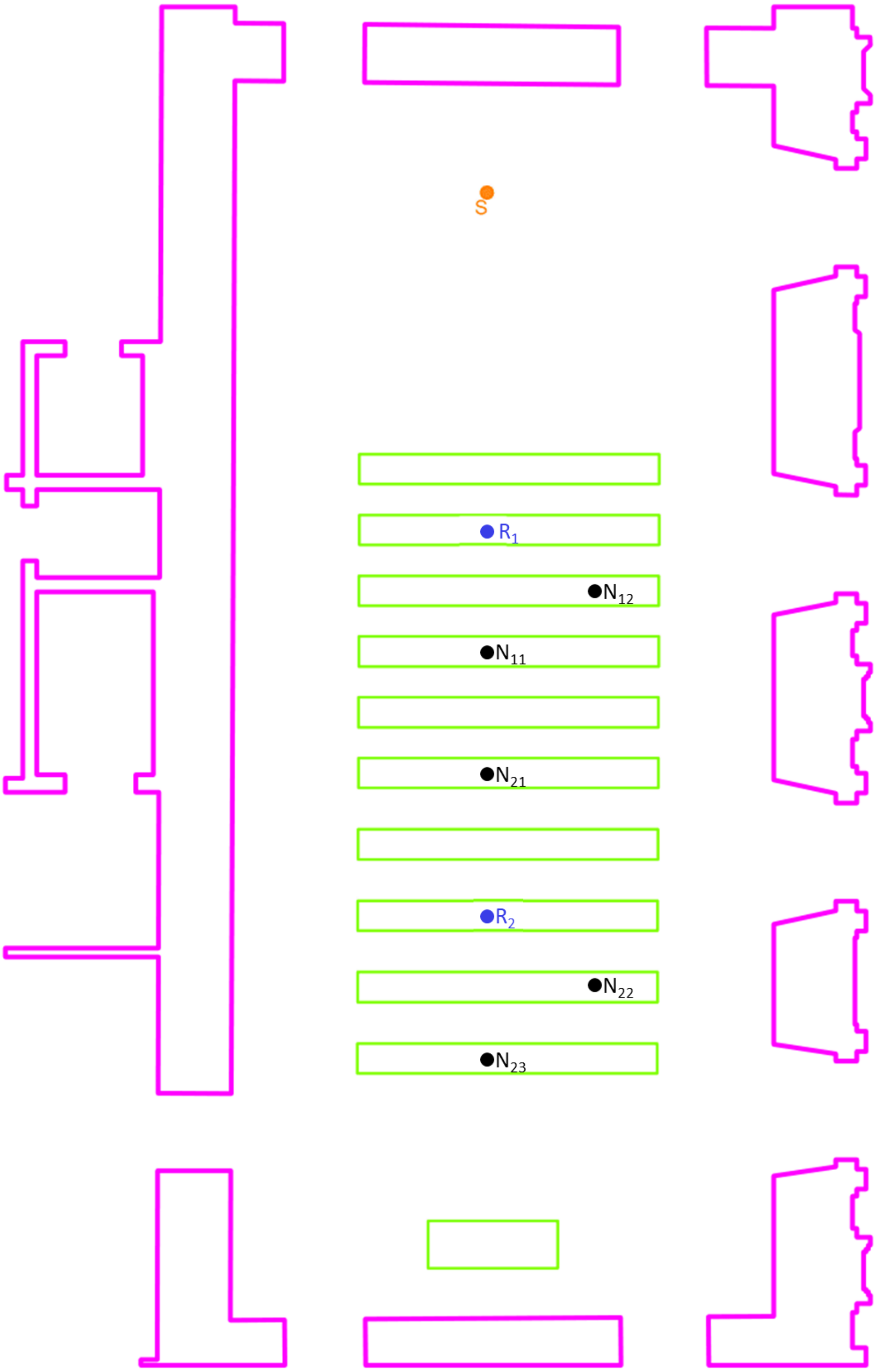


Figure 3.1: The disposition of source, receivers and noise sources in the scenes of the conference hall.

Receiver	Angle LS1	Distance LS1	Angle LS2	Distance LS2
R_1	-67°	6.7 m	62°	5.6 m
R_2	-49°	8.2 m	43°	7.3 m

Table 3.3: The relative arrangement of the loudspeakers to the two reception points in the classroom.

The number of video footage is double the number of the scenes because the noise was visually represented in two different ways filmed separately: the dodecahedron itself used for the RIRs and a small group of four people acting to converse around the point of the noise source. In the recordings without noise sources, the extra people present was randomly sitting in the classroom and simulated to normally follow a lesson.

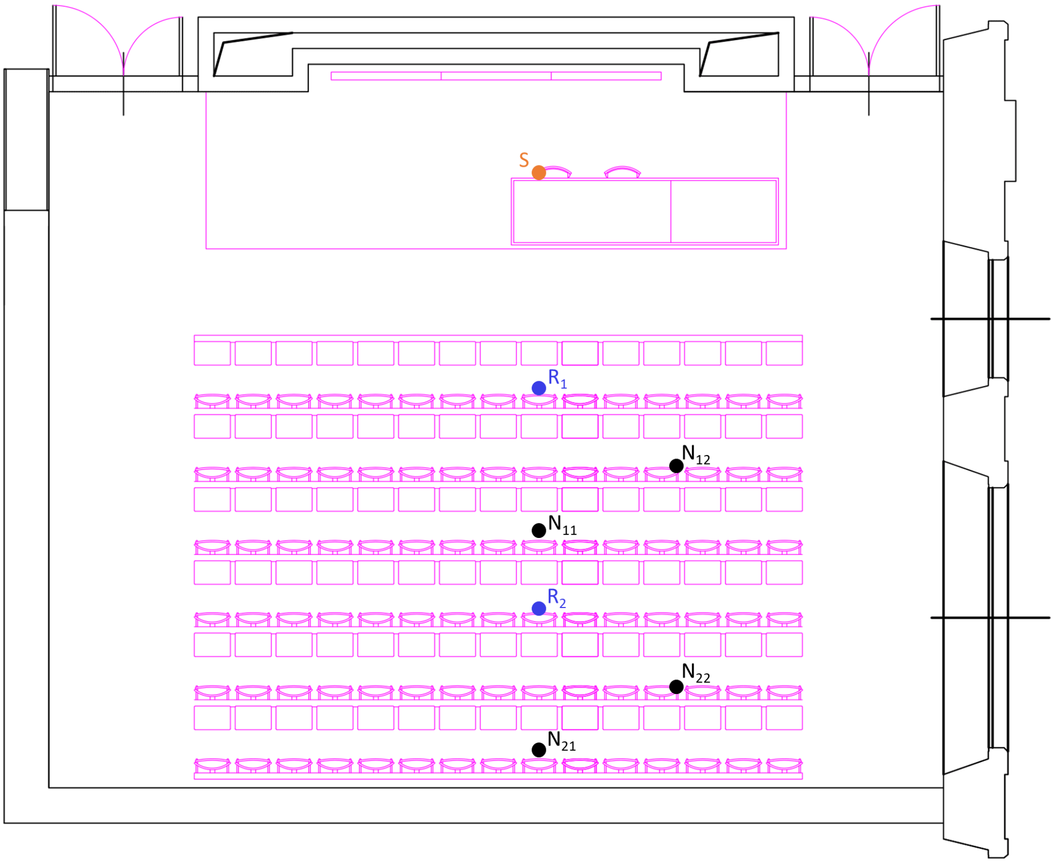


Figure 3.2: The disposition of source, receivers and noise sources in the scenes of the classroom.

3.2 Equipment for the scenes recording

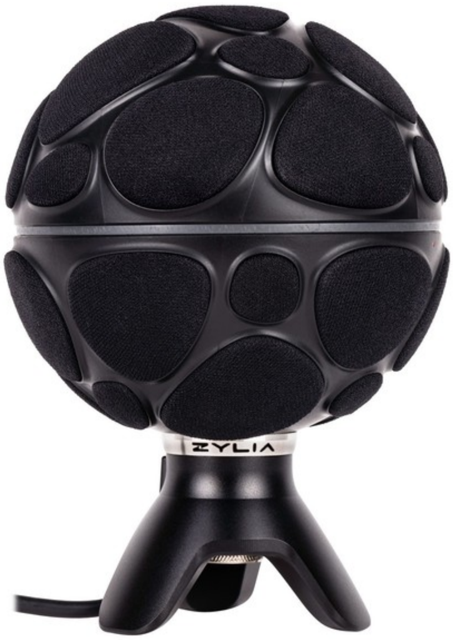


Figure 3.3: The Zylia ZM-1 microphone.

The microphone used to make ambient acoustic recordings and to extract RIRs in Ambisonics is the Zylia ZM-1, which consists of a spherical microphone array. The main technical parameters are mentioned in the table 3.4.

Thanks to its technical characteristics, this microphone can record with great fidelity and without distortion in wide sound fields and over a range of frequencies that comprehends a great variety of sounds. The components that compose it are all digital, and this feature allows to avoid the creation of further noise after acquisition and maintain a high SNR. Furthermore, the 19 capsules have acoustic parameters that are equal to each other (with an uncertainty of +/- 0.1 dB between the different capsules) and constant over time, and thanks to this the microphone never needs to be re-calibrated.

The recordings are not made directly in Ambisonics: the microphone returns the 19 tracks of its capsules, which are subsequently treated with the appropriate filters to obtain the 16 tracks of the 3^{rd} order B-format Ambisonics.

This microphone is very convenient because, by connecting it to a PC with a USB cable, it is possible to record using Zylia's dedicated software or any other DAW. It is therefore not necessary to have additional external components to make the recordings. It has also a LED ring that changes color to indicate recording

Number of capsules	19
Dimensions	Diameter: 103 mm Height: 108.6 mm, 155 mm with stand
Weight	470 g (with stand)
Type of capsules	Omnidirectional XENSIV digital MEMS
Sampling rate	48 kHz
Bit depth	24 bit
Frequency range	20 Hz - 20 kHz
Max SPL	130 dB SPL
SNR	69 dB(A)
Dynamic range	105 dB
THD	<1% up to 128dB SPL
Noise floor	-105 dBFS(A)
Max Ambisonics order	3
Connector	microUSB 2.0
Storage	On external device (e.g. PC)

Table 3.4: Properties of Zylia ZM-1

status: blue when the microphone is connected, and red when it's recording.

The two cameras available, both of the Insta360 brand, have radically different characteristics. The first one, the ONE X2 model, is an action cam. It is therefore optimized for shooting in critical conditions where low weight and agile handling are the priorities (e.g. skiing or motocross). The size and weight are extremely small, as also the physical size of the battery and the storage support (micro-SD). The components for recording are reduced to the minimum necessary to have a 360 shooting, so there are only two fisheyes on the two largest faces of the parallelepiped, with the center placed on the same axis and oriented towards each other in opposite directions. It is a consumer-type device, therefore projected for users who do not have professional needs for the quality of the videos, and also its price is in line with this type of use.

The second video camera available is the model called "Pro". It is thought for professional users, mainly for shooting videos for non-amateur products such as short films in 360 or clips shooted to be included in multimedia projects (e.g. video games, installations). The weight and dimensions are considerably greater than ONE X2 because the use to which the Pro model is dedicated does not consider

Figure 3.4: Insta360 ONE X2

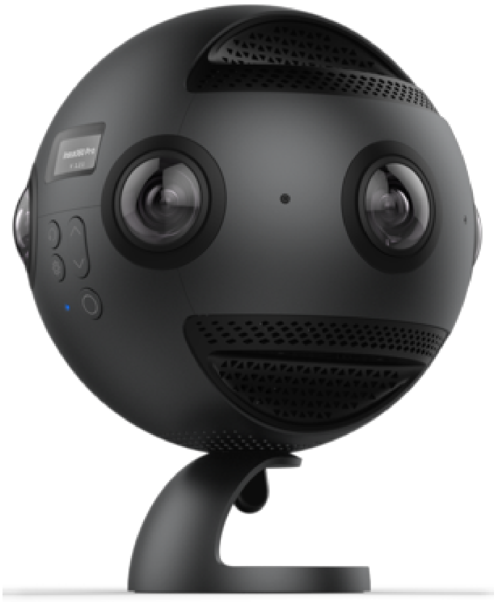


Figure 3.5: Insta360 Pro

agility and portability as requirements. The optimization of this device is aimed exclusively at the quality of the videos, which are mainly made with a still point of view or with a few slow camera movements. The shape is spherical, and the 6 fisheyes are placed on the equatorial circumference. Also for this device, the level of the price is related to the level of its professionalism. The diversity of needs also allows using, in this camera, a standard-size SD card.

The maximum resolution of the Pro is 8K, higher than that of the ONE X2 which is 5.7K for videos and 6K for photos. The ONE X2 is better for the framerate, which can reach up to 60 fps against the 30 fps of the Pro. This difference is given by the fact that being the ONE series of Insta360 focused on the action-cam sector, the shots of these cameras must be fluid to have more continuity of movement because the camera makes often fast movements. The high frame rate shots of the ONE X2 are also adaptable to do slow-motion clips.

The battery of the ONE X2 is optimized to have a good compromise between size and charge duration. That one of the Pro, despite being much more capable, has a shorter duration of use caused by the high energy consumed by the fisheyes (the quantity of them is triple compared to that one of the ONE X2). The shorter duration is not seen as a defect, because in work environments where a video camera like the Pro is used, there is the habit of keeping other batteries in charge during the use of the camera, and they are changed several times during the shooting session. To contrast the heat caused by the high amount of energy dissipated by the Insta360

	ONE X2	Pro
Dimensions	46.2 x 113 x 29.8 mm	Diameter: 143 mm
Weight	149 g	1228 g
Number of fisheyes	2	6
Type of lenses	7.2 mm F2	200° F2.4
Storage	Micro SD UHS-I V30	SD V30 or above USB 3.0 Hard Disk
Interface	USB Type-C Bluetooth BLE 4.2 Wi-Fi	HDMI 2.0 Type-D RJ45 Ethernet interface USB Type-C WiFi
Max 360 photos resolution	6080 x 3040	8K (7680x3840)
Max 360 videos resolution	5.7K 30fps 4K 50fps 3K 100fps	8K (7680x3840) 30fps
Max 360 3D photos resolution	No 3D available	8K (7680x7680)
Max 360 3D videos resolution	No 3D available	6K (6400x6400) 30fps
Video coding	H.264/H.265	H.264
Battery capacity	1630 mAh	5100 mAh
Battery autonomy	80 min recording at 5.7K 30fps	75 minutes not specified the use
Recharging type	USB-C cable	USB-C cable external charger
List price at release	429.99 $	3499 $
Year of release	2020	2017

Table 3.5: Comparing properties of Insta360 ONE X2 and Insta360 Pro

Pro, this camcorder contains a cooling fan system that stays on continuously when the camcorder is turned on. A firmware update allows automatically turning them off only when it is filming a video, to not maintain the fan noise in the recordings.

Since the Pro is mainly used in stationary positions during recordings, and the video files that it records are generally larger (about 31 Mbps bitrate), it is also possible to record directly on an external Hard Disk.

A final advantage that the Pro offers is the possibility of shooting in 3D, at a maximum resolution of 6K for both photos and videos.

Comparing the two cameras, the conclusion is that the Insta360 Pro is much more appropriate to the needs of the shooting of this project than the ONE X2. The need for realism leads to the requirement to have the highest possible resolution, and also 3D can help to make the vision as ecological as possible. The scenes that will be filmed have a fixed point of view, so are not needed the handling and agility that the physical device of the ONE X2 allows sacrificing the video quality and the definition.

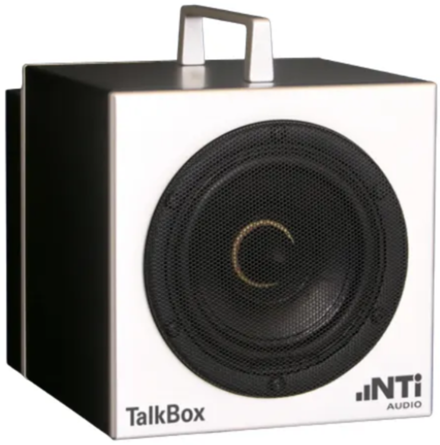


Figure 3.6: Nti Talkbox

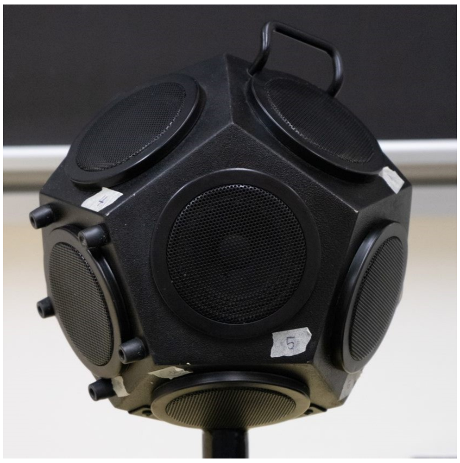


Figure 3.7: Bruel&Kjær 4292-L

Two different types of sources were used as the output of the sound signals used for impulse response measurements. The first was the TalkBox, an emitter of sound signals produced by NTi which has, as its most important feature, a sound emission directivity similar to human speech. Sweeps played with the TalkBox are loaded onto a memory card which is inserted into the TalkBox itself. It, therefore, does not need to be connected to other external devices to emit the signals. The second type of emitter is an omnidirectional dodecahedron (model 4292-L by Brüel&Kjær). This, to play sounds, needs to be connected to an amplifier. In this case, the amplifier where the gain was controlled (model LAB300 by Lab Gruppen) had a Tascam US144 sound card as input. This card was connected to a computer, from which the tracks were played using the DAW Audition of Adobe. Another object used was the NTi XL2 sound level meter (SLM), used in the measurements because it offers also the function of an omnidirectional microphone and recorder. The last tool to mention is Bidule: it is a DAW that has a block-based structure and, in this case, was used to record the 19 microphone tracks of the Zylia and to convert them into 3rd-order Ambisonics.

3.3 In-field measurements procedure

Before describing the measurements performed in the environments, it is necessary to make some clarifications that will help to contextualize the description of the measurements workflow.

Throughout both measurement sessions, was maintained a height of 1.20m for the centers of the acoustic and visual emitters and receivers. According to the referenced standard, ISO 3382-1 and ISO 3382-2, acoustic measurements have to be done in an empty environment. In the case of the measurements described here, there were two people inside the room; this situation, according to the standard, can still be considered an empty room. The sweep used was a sine sweep with progressive frequency from 20 Hz to 20 kHz, three repetitions of 5 s interspersed with 5 s of silence, fade-in and fade-out of 0.1 s, and a halved dynamic. All recordings made with the Zylia microphone were saved, using the Bidule software, in three different formats:

- 19-ch (RAW registrations of microphone capsules);
- ambix (3rd-order Ambisonics, 16 channels);
- W (mono-channel omnidirectional track, corresponding at the 0-order Ambisonics).

The SLM was used exclusively as an omnidirectional recorder for this session.

Before proceeding with the measurements, a few preparation steps had to be performed. The first thing done was to calibrate the SLM, making it record a signal with known frequency and SPL (1000 Hz and 94 dB) emitted by a calibrator. With the same SLM, immediately afterward, the environmental noise was recorded for 5 minutes in a single reception point. Subsequently, a first measurement of the sweep emitted by the TalkBox was made with the SLM. This measurement was used to calculate the SNR by octave bands using, as reference noise, the ambient noise just recorded. This SNR is useful to understand if the reverberation time measurement will be reliable, and must be at least 45 dB to calculate the T_{30} and 35 dB to calculate the T_{20}. Once this verification was performed with MATLAB, it is possible to proceed with the measurement.

The first phase consisted in recording some IRs which would be used to calculate the acoustic parameters of the environment. 5 random points were defined to do recordings there and, for each of them, the sweep was emitted six times by the TalkBox: three times it was recorded by the SLM and three times by the Zylia.

During these first measurements, in every position in which the Zylia was placed, it was always oriented keeping its frontal direction towards the TalkBox. The acoustic data to be extrapolated from these measurements was calculated in post-processing at another moment.

The second part of the audio measurements was done to calculate the RIRs for the target and noise sources from listener positions. In the two receiving positions defined, R_1 and R_2, the sweeps emitted by the TalkBox in the target position T were recorded by the Zylia (see figure 3.8). The measurements to obtain the RIRs of the target signal were performed twice: firstly emitting the sweep exclusively from the TalkBox, and secondly by positioning the microphone in front of the TalkBox so that the signal was amplified by the room's sound system.

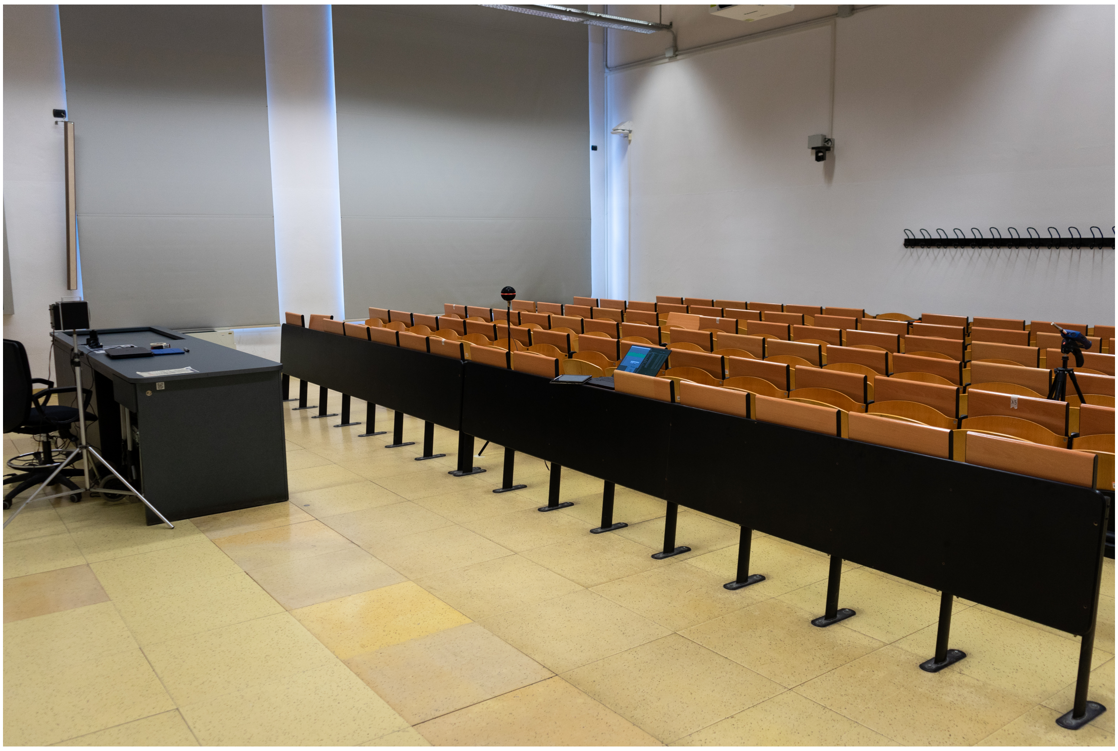


Figure 3.8: Measuring a RIR associated with the target source in the classroom.

Subsequently, for each reception position R, the sweeps coming from the noise source placed in the masking noise positions M decided for the specific listening positions were recorded by the Zylia (see figure 3.9). The emitter to measure the RIRs in the noise positions was used, respectively, the TalkBox for the "one-talker" type directional noise (the case of the conference hall) and the dodecahedron for the omnidirectional SSN noise (the case of the classroom).

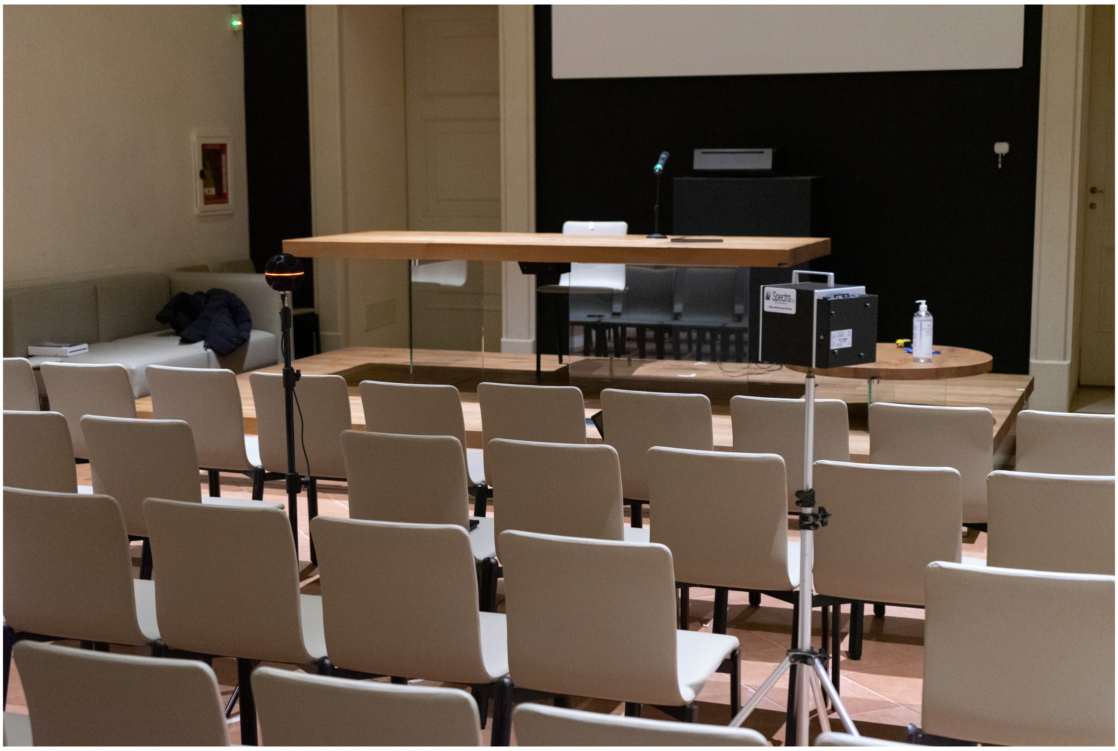


Figure 3.9: Measuring a RIR associated with a noise source in the conference hall.

The third part of the measurements was for the 360 videos. The workflow to shoot with Insta360 Pro was done as follows. The battery inserted in the camera must be full of charge, because the consumption is high when the camera is on, and the power on takes about 2 minutes. When is the first time shooting with a formatted SD, must be done the speed test of the SD, which requires about 4-5 minutes. Every time the camera placement is changed, the process of calibration must be done to have a better stitching quality later. Finally, to begin and stop the recording, is better to use the smartphone app “Insta360 Pro” to avoid the presence of a person near the camera who press the recording button at the beginning and the end of the video. Before starting, it must be decided the type of the shot, and in this case, was done in 3D 6K to have the possibility to obtain both 6K 2D and 3D footage at the same time. The video camera was placed in the place of the Zylia at receiver points. For each defined scene, the visual recordings were made both with and without a microphone in front of the TalkBox used as the source.

3.4 Room Acoustic Characterization

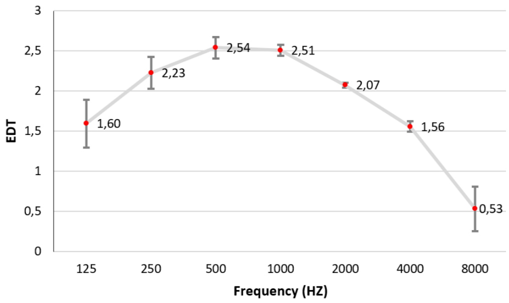
The RIRs are obtained mathematically in this way starting from the emitted sweep and the recording of the sweep made by the microphone:

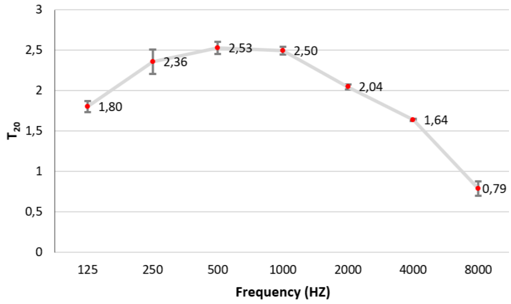
$$Rec(t) = h(t) * Sweep(t)$$
$$Rec(f) = h(f) \cdot Sweep(f)$$
$$h(f) = \frac{Rec(f)}{Sweep(f)} = Rec(f) \cdot Sweep(f)^{-1}$$
$$h(t) = Rec(t) * Sweep(t)^{-1}$$

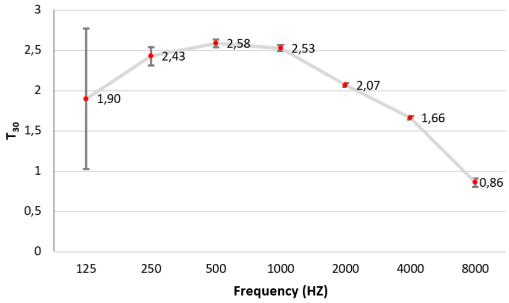
The final expression obtained is the one to calculate the impulse response $h(t)$ from the initial sweep $Sweep(t)$ and its version recorded by the microphone at the reception point $Rec(t)$.

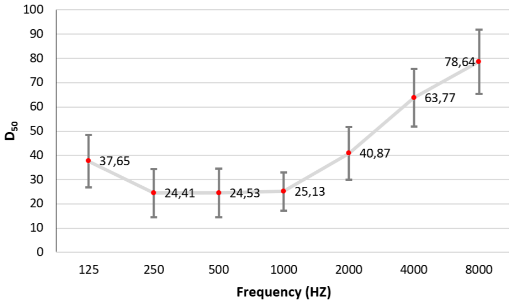
These are the results of the calculations described in the section 2.1.2 to obtain the acoustic parameters by octave bands of the two rooms.

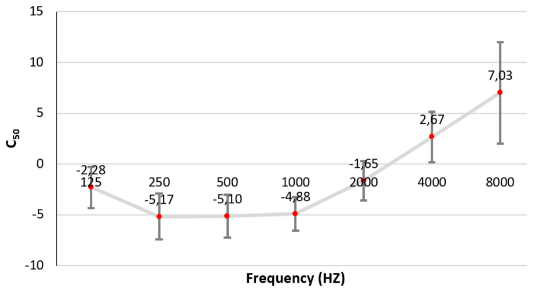
Results for the classroom:

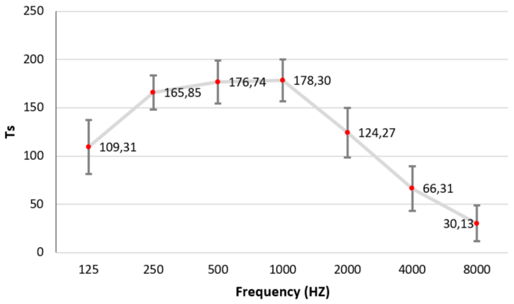




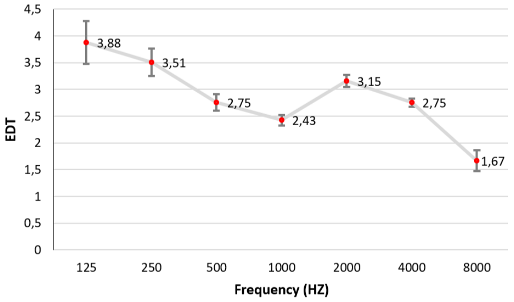


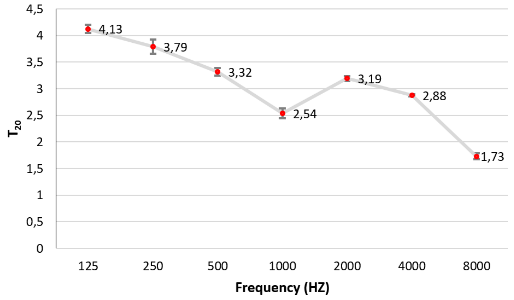


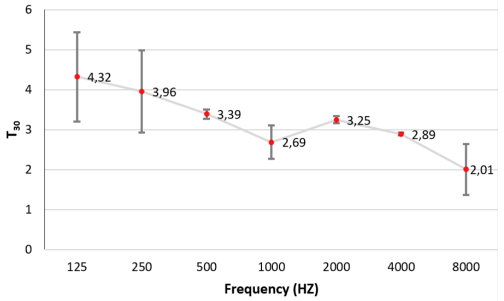


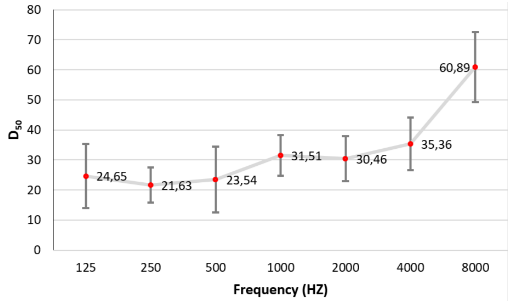


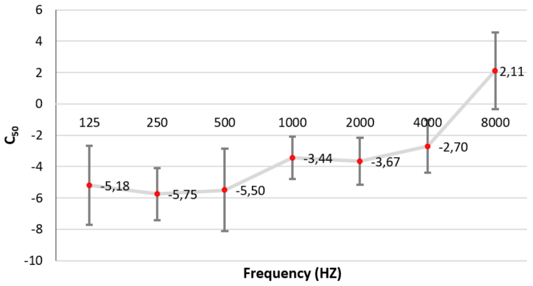
Results for the conference hall:

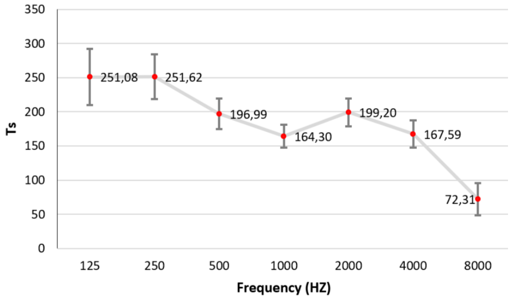












3.5 Video post-processing procedure

The process that is about to be described makes it possible to create files for the immersive fruition with the Oculus Quest 2 viewer starting from the files exported from the memory cards of the two camcorders. The two software programs used to open these files are different, depending on the model of the camera, and are respectively called "Insta360 Studio" for the ONE X2 camera and "Insta360 Stitcher" for the Pro model. These two programs are both free for the owners of one of the camera models compatible with them. Their main purpose is "stitching," which is the process of automatically joining the footage taken by the different fisheyes. The stitching ends with the creation of a single flat 2D file that can be put into any editing program to do post-production of videos. This file is called "equirectangular" (figure 3.10), has a 2:1 aspect ratio, and contains all the visual content present in a 360 video, extracted from the spherical surface where it is ideally located doing an equirectangular projection.

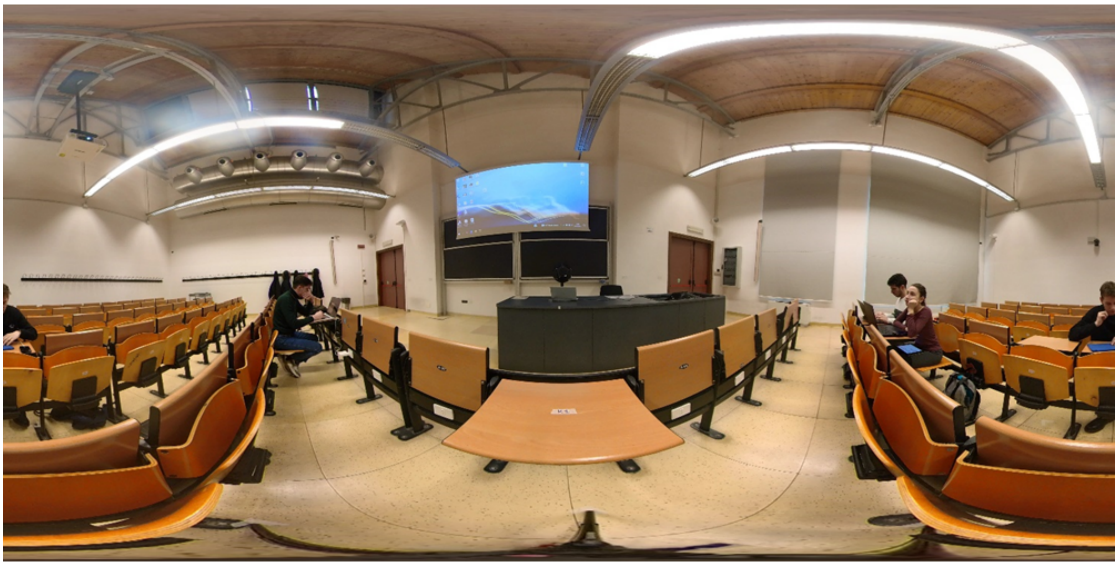
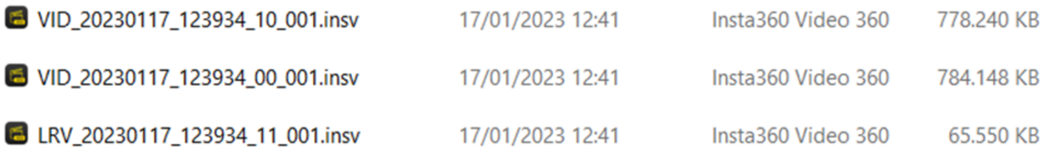


Figure 3.10: An example of an equirectangular video frame of the session of recordings made in the classroom.

In the SD card of the ONE X2, videos in 360 are saved in three different files (figure 3.11).



VID_20230117_123934_10_001.insv	17/01/2023 12:41	Insta360 Video 360	778.240 KB
VID_20230117_123934_00_001.insv	17/01/2023 12:41	Insta360 Video 360	784.148 KB
LRV_20230117_123934_11_001.insv	17/01/2023 12:41	Insta360 Video 360	65.550 KB

Figure 3.11: The three different files for a single 360 video in ONE X2.

They, which are in ".insv" (Insta360 Video 360) format, contain, respectively, the separate footage created by the two different fisheyes at maximum resolution (files with the name beginning with "VID") and a low-resolution version optimized for preview playback directly from the camera display or from the smartphone that controls it (files with the name beginning with "LRV," i.e., "Low Video Resolution").

The elements that identify groups of files made for the same recording are present in the name of the files themselves: date, time, and the final enumeration. The penultimate field contains three different identifiers (10, 00, 11) to differentiate the three files from each other.

On any computer with Insta360 Studio software installed, opening arbitrarily one of these three files will automatically open the program with the full-resolution videos of the selected shot loaded (figure 3.12).

In the program, the stitching is done automatically and it is possible to navigate inside the 360 video before starting the desired type of export. The main function that this program offers is to reframe, which means to create flat versions of the video that include only a part of the spherical portion (is possible to move the window of view during the duration of the video), but this is not the function to use this project.

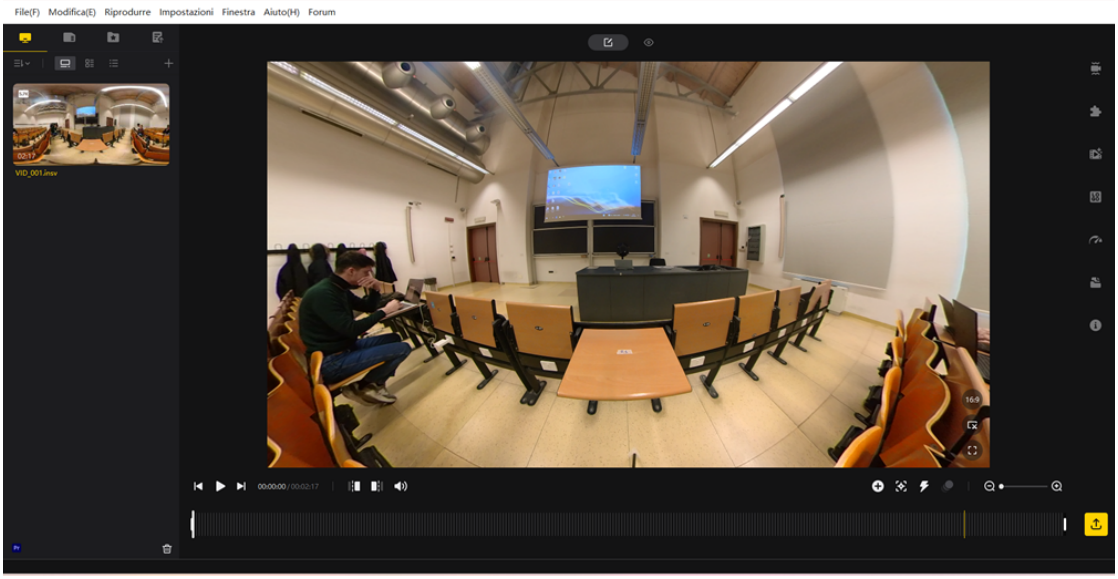


Figure 3.12: The main interface of the software Insta360 Studio.

For the aim of this project, is sufficient to take a brief check at the video to verify the quality of the stitching and, in case, to crop the duration in the timeline if there are visible movements of the beginning or end of the shots (e.g. in some of our scenes the characters start to simulate the scene a few moments after the start of the recording or stop doing it a few moments before the end). To do this, the user must navigate with the mouse by clicking and dragging the field of view inside the viewing window and, in case of a need for cuts, move the white indicators at the beginning and end of the timeline to delete some frames.

In the vertical bar on the right, some settings must be checked before starting the export. For in this case, the optimal settings will be as follows, divided into the sections in which they are located in the program.

- Stabilization Type
  - FlowState Stabilization ON: it serves especially to stabilize moving shots, but also to contain the oscillatory micro-movements that the video camera can have during still shots like ours.
  - Direction Lock OFF: it is used to keep the direction of vision constant in moving shots in which the direction of vision of the camera is changed during the registration. For the shooting done here, it is completely useless.
- Stitching
  - Stitching set to NORMAL: this is the right option to use if there are no additional lens caps or cases such as the underwater one.
  - Dynamic Stitching OFF: Time-variable stitching, ideal for scenes with many rapid movements different from ours.
  - Optical Flow Stitching ON: constant stitching for each frame, but done with a more accurate and precise algorithm than the "Dynamic Stitching". It is ideal for the scenes made, with little movement and a strong need for realism.
  - Chromatic Calibration ON: it is a function that uniforms the color differences perceived by the different fisheyes, to create a color continuity in the stitching areas.
  - Stitching Calibration: serves to recalibrate the stitching and is to be used only in case of continuity imperfections visible in the junction area of the images recorded by the different fisheyes.
- The other sections (Media Processing. Logo, Stats, Project Management) contain various settings that are not of interest to this project and all the options contained in them must be kept deactivated.

At this point, it is possible to proceed with the export by clicking the "Start export" command with the yellow button at the right of the timeline. The box that opens is the one in figure 3.13. It is fundamental, in the first line, to select the "Export 360 video" option, which allows exporting the equirectangular video. In addition to changing the file's name and the destination folder, the most influential choice is the selection of the codec "Encoding Format". The proposals we have in this selection are the following.

- H.264 or H.265

  If the work needs to stop here with the post-production workflow and keep the video as-is to view in the HMD, can be selected one of these first two choices. They are lossy codecs optimized to don't compromise the quality of vision and to obtain the smallest possible file size. In particular, H.265 is the more modern of the two, which allows for having, at equivalent quality, a much smaller file size than H.264. For both codecs, an export bitrate can be decided, but the one preset by the program (90 Mbps) is already a good compromise between perceptual quality and file size. In the case of the current project, it will be necessary to have a not excessively compressed file that can allow it to be further post-produced in video-editing software to eliminate the visible tripod that holds the video camera.

- ProRes 422

  This is a compressed codec, so returns a file of similar quality to that one of an uncompressed codec. The file size will be very higher, but suitable for future steps of post-production that, with codecs such as H.264 or H.265, would cause excessive degradation of video quality. Furthermore, less heavy compression allows having a file that can be handled in a computationally lighter way by video editing software. Finally, with Apple ProRes, the bitrate is determined exclusively by the codec and cannot be changed. For the type of work to do for this study, is preferable to export with it.

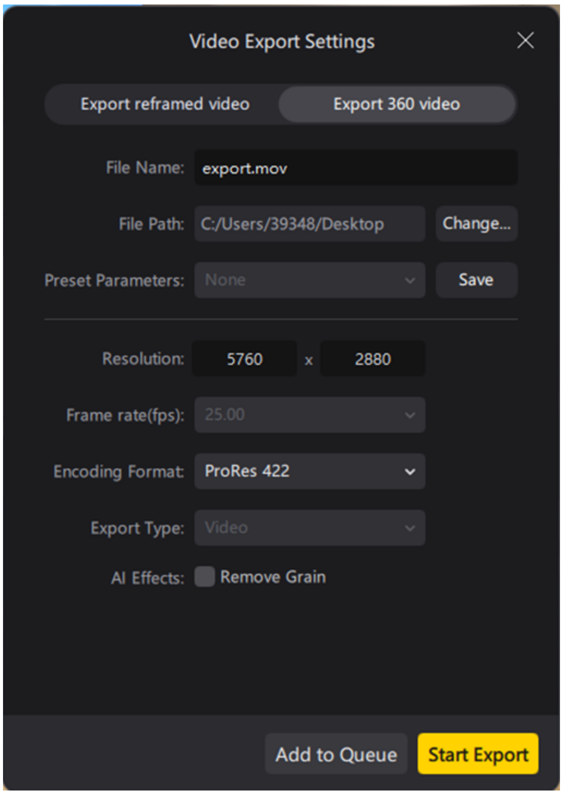
The other parameters, such as resolution and framerate, are already preset by the program according to the resolution of the file and must not be changed.

To compare the file sizes for the various codecs, different exports were performed with 20 seconds of test footage, and the resulting file sizes were as follows (at 25 fps and 5760 x 2880 resolution):

- H.264/H.265 at 90 Mbps $\sim$ 218 MB
- Apple ProRes 422 $\sim$ 2,22 GB

The last checking option, "Remove Grain", allows to partially limitate the disturbing grain that is typically present in videos done in low light or artificial lighting conditions in interiors. In this case, it is advisable to keep it selected to have a slight qualitative improvement.

After explaining the process of obtaining the equirectangular video for the Insta360 ONE X2, it is necessary to explain how to get to the same point with the material shot by the Insta360 Pro camera.



Video Export Settings

File Name: export.mov

File Path: C:/Users/39348/Desktop

Preset Parameters: None

Resolution: 5760 x 2880

Frame rate(fps): 25.00

Encoding Format: ProRes 422

Export Type: Video

AI Effects: □ Remove Grain

Figure 3.13: The box of export settings of Insta360 Studio.

Videos shot with the Pro model are saved in separate folders dedicated to the single video. The name of the folder begins with "VID", followed by the fields containing the date and the hour of the shooting start time. Inside the folder, the content appears as shown in figure 3.14.

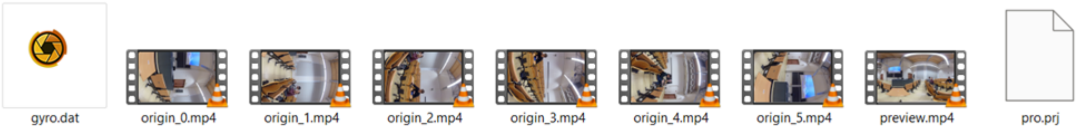
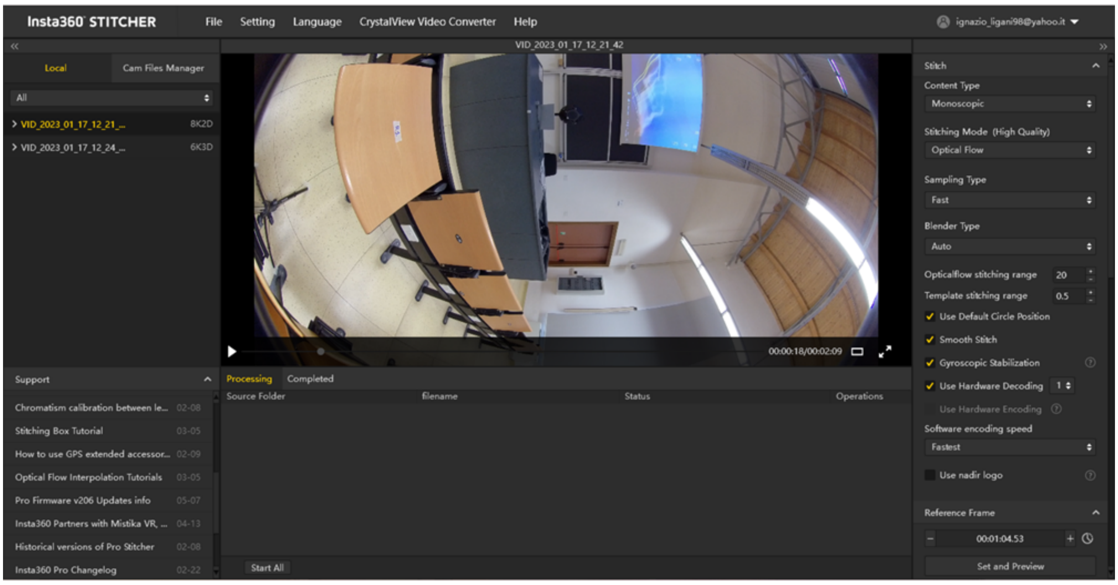


Figure 3.14: The different files made for a 360 video with the Pro model.

In addition to files with metadata and proxies, there are separately the 6 mp4 videos recorded by the different fisheyes.

Insta360 Stitcher (figure 3.15) is a less user-friendly program than Insta360 Studio because it is made for a more professional user with more technical knowledge. To import the video into the program, the entire folder must be dragged and dropped into the program's file list area.


Insta360 STITCHER
File
Setting
Language
CrystalView Video Converter
Help
ignazio_ligani98@yahoo.it
VID_2023_01_17_12_21_42
Local
Cam Files Manager
All
VID_2023_01_17_12_21_... 8K2D
VID_2023_01_17_12_24_... 6K3D
Support
Chromatism calibration between le... 02-08
Stitching Box Tutorial 03-05
How to use GPS extended accessor... 02-09
Optical Flow Interpolation Tutorials 03-05
Pro Firmware v206 Updates info 05-07
Insta360 Partners with Mistika VR, ... 04-13
Historical versions of Pro Stitcher 02-08
Insta360 Pro Changelog 02-22
00:00:18/00:02:09
Processing
Completed
Source Folder
filename
Status
Operations
Start All
Stitch
Reference Frame
Set and Preview

Content Type: Monoscopic

Stitching Mode (High Quality): Optical Flow

Sampling Type: Fast

Blender Type: Auto

Opticalflow stitching range: 20

Template stitching range: 0.5

☑ Use Default Circle Position

☑ Smooth Stitch

☑ Gyroscopic Stabilization

☑ Use Hardware Decoding 1

□ Use Hardware Encoding

Software encoding speed: Fastest

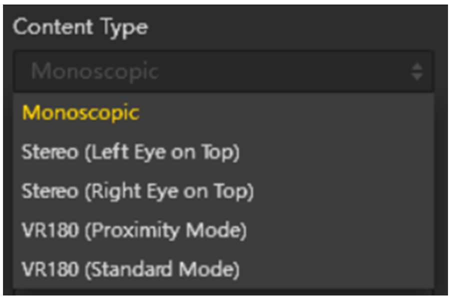
□ Use nadir logo

Reference Frame: 00:01:04.53

Figure 3.15: The interface of Insta360 Stitcher with 2 videos loaded.

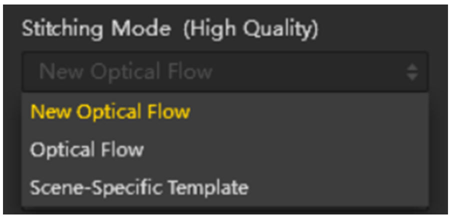
In this program, the user can use different stitching algorithms to find the better one for the captured footage. Often several trials have to be made by changing the parameters until the optimal combination is found. The parameters that must be checked are found in the bar on the right, and here the different options will be described concording with how are explained in the official online manual on Insta360 website [30].

- Content Type



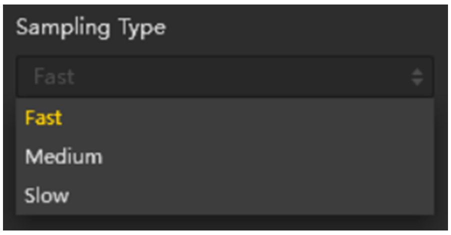
In this box can be selected the type of content to export. It must be specified that the monoscopic shots made with maximum resolution (8K) can only be exported with the "Monoscopic" option. The stereoscopic ones, with a maximum resolution of 6K, can also be adapted as monoscopic, maintaining the reduced 6K resolution. For stereoscopic export, must be selected the first option "Left Eye on Top" because it is universally recognized by reproduction systems like Oculus HMDs.

- Stitching Mode



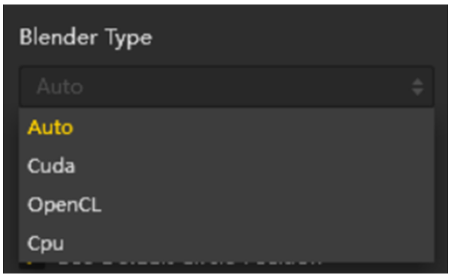
For this selection, it is highly recommended to decide on one of the first two options. They are similar, but "New Optical Flow" offers stitching speeds up to three times faster than "Optical Flow", with the disadvantage that the second often offers greater stitching accuracy than the first. The recommended choice is therefore to try the first one and, if not satisfied, try the second one. It is not necessary to render the whole video to test, is possible to test with just a frame using a command that will be analyzed later in the explanation. “Scene-Specific Template” is a non-optical flow algorithm and therefore extremely unreliable for stitching on indoor shots.

- Sampling Type



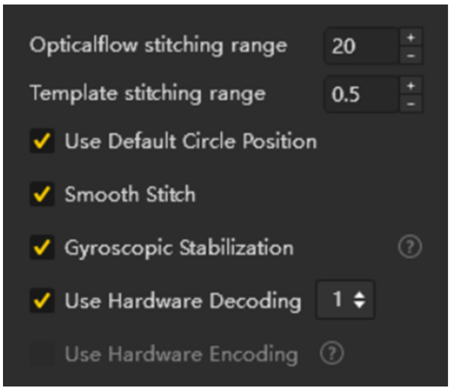
Also for this setting, some changes can be tried just if the first stitching trials are not satisfying. Slower sampling generally leads to more accurate stitching results. The difference in results is especially evident for videos taken with the camera in movement.

- Blender Type



This choice is used to decide which type of hardware acceleration to use. Here too it is advisable to keep the automatic choice, which will set the optimal algorithm according to the hardware possessed by the machine: "Cuda" for Nvidia graphics cards, "OpenCL" for non-Nvidia graphics cards, and "Cpu" in case it is not possible to do anything other than run the calculations directly into the processor. In this last case, sometimes the rendering does not start automatically with "Auto" (an error is reported), and if this happens we have to manually select "Cpu".

- Other stitching options


Opticalflow stitching range 20

Template stitching range 0.5

✓ Use Default Circle Position

✓ Smooth Stitch

✓ Gyroscopic Stabilization

✓ Use Hardware Decoding 1

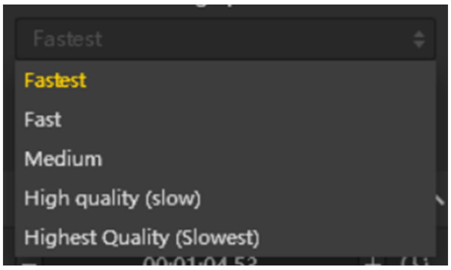
□ Use Hardware Encoding

To have good stitching must be selected all of the check box possible (depending on our hardware) here.

“Opticalflow stitch range” is a number related to the width of the image area considered in the optical flow conjunction, and should be kept at the maximum possible value: 20 for 2D shooting and 16 for 3D shooting.

“Template stitching range” is related to the width of the stitching line, and the value of 0.5 is ideal for having a good compromise between the smoothness of the passage and the prevention of the ghosting effect.

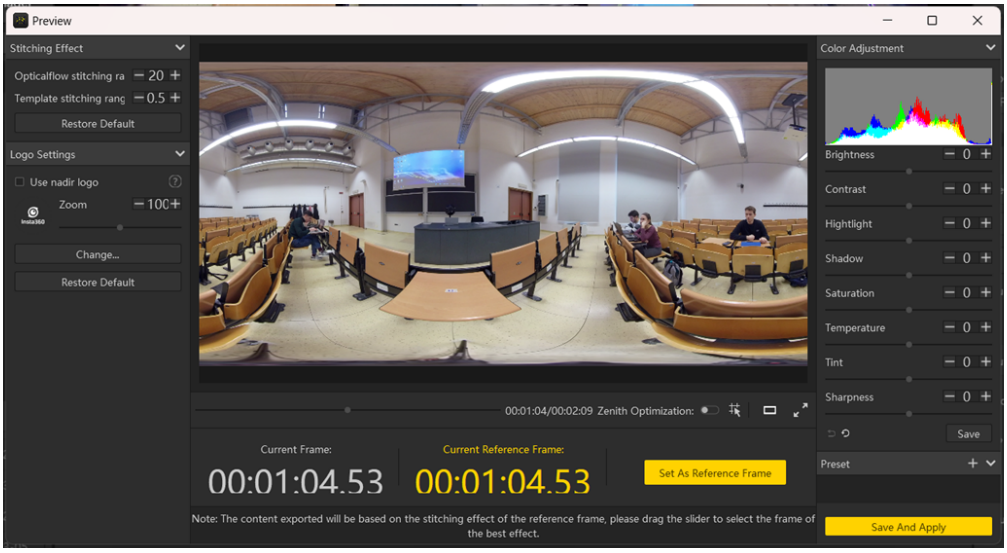
- Software encoding speed



Also for this selection, it is preferable to start from the pre-set "Fastest" setting and then go to decrease the encoding speed in case the test results are not satisfactory. However, it must be considered that the selection made here brings differences only in the case of software encoding ("Cpu" selection in Blender Type).

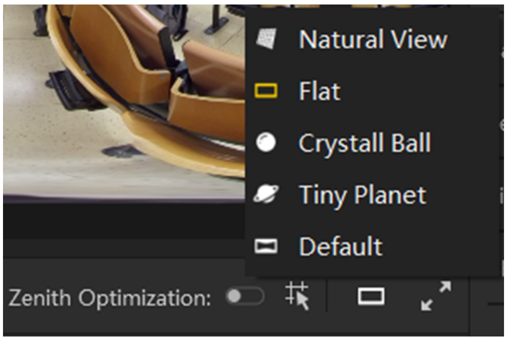
- Reference Frame

  This is the most important section to consider for efficacious stitching. After selecting a time instant of the video to be used as a reference, which can be chosen arbitrarily without problems, clicking on "Set and Preview" will open a window with a preview of the stitching rendering for the selected frame.



Opticalflow stitching ra 20
Template stitching rang 0.5
□ Use nadir logo
Zoom 100
Brightness 0
Contrast 0
Hightlight 0
Shadow 0
Saturation 0
Temperature 0
Tint 0
Sharpness 0
00:01:04/00:02:09 Zenith Optimization:
Current Frame: 00:01:04.53
Current Reference Frame: 00:01:04.53
Note: The content exported will be based on the stitching effect of the reference frame, please drag the slider to select the frame of the best effect.

In addition to seeing the equirectangular, there is the possibility to select other viewing modes.



Among these, a very useful one is the "Natural View". Using this, it is possible to navigate the spherical surface of the video in the same way as already seen on the main screen of the Insta360 Studio program.

In this box, other two relevant things can be done. The first is the Zenith Optimization, which offers a slight improvement in scenes containing, mainly above or below the camera, regular line patterns (for example square tiles or ventilation system ducts). The second is the reorienting of the 360 shot, to re-center the initial direction of view of the video. However, it is not recommended to do this here, because there is no precise control over the angles of movement on the three axes, and there is the possibility to correct the centering just by moving the shot with the mouse. This correction will be done more precisely after, in the video-editing program.

The other settings offered are mainly related to the color correction of the image but are not of interest to the purpose of this work. If the stitching preview is satisfactory, should be clicked “Set As Reference Frame” and “Save and Apply”. If not, it must be closed the panel, changed the previous parameters, and repeat the preview until an acceptable result is found.

The remaining commands are to set the start and end frame between which the user wants to export (equivalent to the white controllers on the Insta360 Studio timeline) and to decide the export codec. Here there are more possibilities than those seen for the previous program, but the ones useful to use, according to the needs of this type of work, are those already seen previously. Has also to be decided the container that determines the file extension, and the recommended ones are “.mp4” for the H.264 and H.265 codecs and “.mov” for the Apple ProRes 422.

For the current project, a .mov file with Apple ProRes 422 HQ codec is the adequate option, at the maximum possible resolution and with “Normal” or “None” audio (we will recreate it, so spatializing the audio recorded by the camera would require an unnecessary use of computational power). The file will be particularly large because the maximum resolution allowed by the Insta360 Pro is significantly higher than that of the camera previously seen.

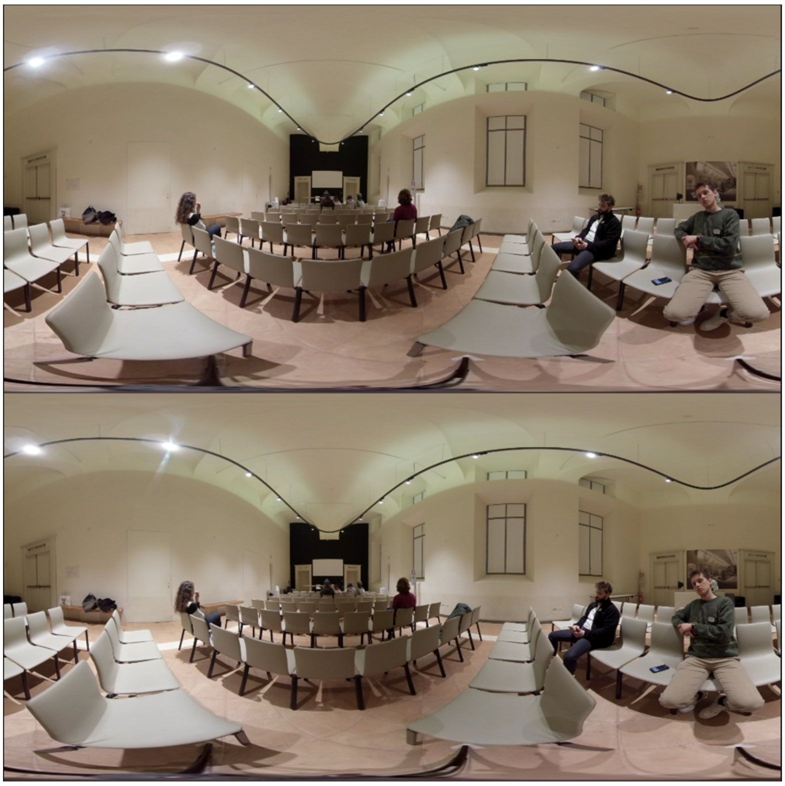
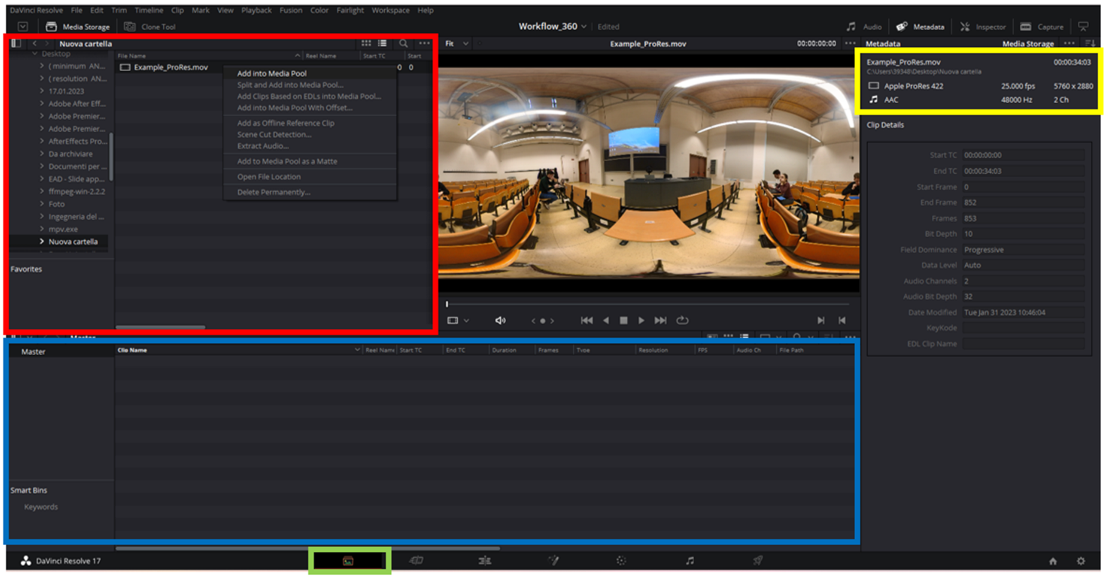


Figure 3.16: An example of equirectangular 3D 360 video frame made at Museo Egizio of Turin.

In the case of 3D videos, two things essentially change: inside this program, it is not allowed to reorient the shots in 3D (it will be necessarily done in the post-production program), and the equirectangular file of a 3D shot (figure 3.16) has a 1:1 proportional aspect, containing the two different 2:1 videos, made for the individual eyes, superimposed.

Now, once created the equirectangular file in Apple ProRes, can be begun the post-production workflow in a video editing program. In this case, it will be done with Blackmagic Design's DaVinci Resolve Studio 17. The most updated version is 18, but the procedure described is the same for both. It is an integrated software with functions of video editing, visual effects, DAW, and color correction. It offers multiple features, but here will be seen just those necessary for our workflow on a 360 video.

The first case analyzed is the one of a 360 2D video. A new project must be created, in which it is not necessary to change the settings of resolution and frame rate because they can be adapted to the current video during the creation of the single timeline. As visible in figure 3.17, firstly must be opened the "Media"



Metadata | Media Storage

Example_ProRes.mov 00:00:34:03
C:\Users\39348\Desktop\Nuova cartella
Apple ProRes 422 25.000 fps 5760 x 2880
AAC 48000 Hz 2 Ch

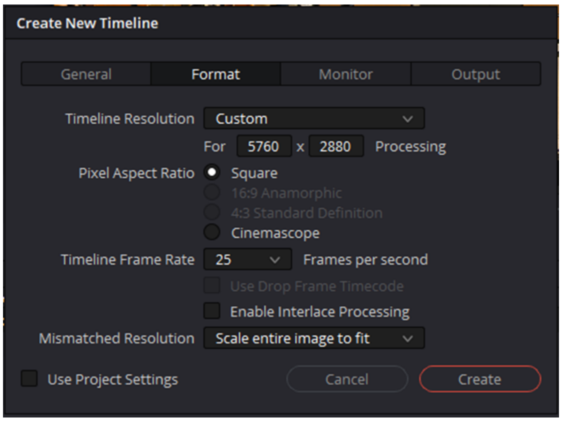
Clip Details

Start TC	00:00:00:00
End TC	00:00:34:03
Start Frame	0
End Frame	852
Frames	853
Bit Depth	10
Field Dominance	Progressive
Data Level	Auto
Audio Channels	2
Audio Bit Depth	32
Date Modified	Tue Jan 31 2023 10:46:04
KeyKode	
EDL Clip Name	

Figure 3.17: The Media section of DaVinci Resolve.

section (indicated in green) and, in the Media Storage, (indicated in red) will be searched the file to import in its path. Once found, we can right-click on it and select “Add into Media Pool”. After doing this, it will be visible in the Media Pool (indicated in blue). A panel will open where it will be proposed to change the project settings depending on the file just imported, but at this moment can be declined the invitation. From now on must be paid attention to the fluidity of playback depending on the power of the computer used. If that’s not fluid enough, will be necessary to create a low-resolution proxy of the footage file by right-clicking on it in the Media Pool and selecting “Generate Optimized Media”. To activate the use on the timeline of the proxies created must be selected, in the upper bar, "Playback -> Use Optimized Media if Available". Another way to make the vision of the video more fluid is to select, in the upper bar, "Playback -> Timeline Proxy Mode -> Half Resolution or Quarter Resolution".

At this point, the timeline can be created by right-clicking on the video in the Media Pool and selecting “Create New Timeline Using Selected Clips”.

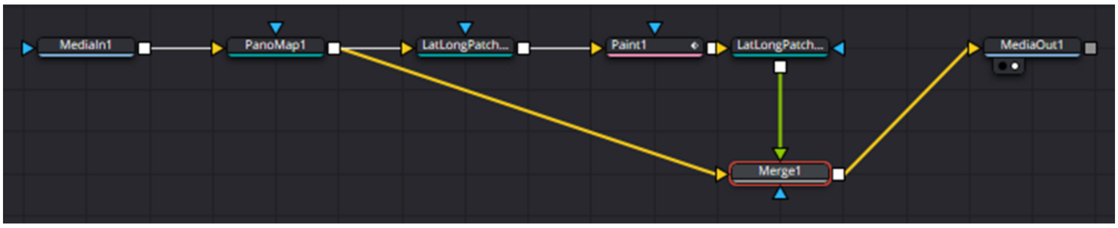

Create New Timeline

Timeline Resolution: Custom
For 5760 x 2880 Processing
Pixel Aspect Ratio: ● Square
○ 16:9 Anamorphic
○ 4:3 Standard Definition
○ Cinemascope
Timeline Frame Rate: 25 Frames per second
□ Use Drop Frame Timecode
□ Enable Interlace Processing
Mismatched Resolution: Scale entire image to fit
□ Use Project Settings

On the panel that opens, must be deselected “Use Project Settings”. In the tab “Format”, must be set the resolution and the framerate according to the information present in the “Metadata” rectangle (yellow in the precedent figure 3.17) that refers to the selected video. In the case of this demonstration, the file is the equirectangular version of a video made with the Insta360 ONE X2 camera, so set the resolution must be set at 5K (5760 x 2880) and the Frame Rate at 25 fps. For the resolution, it is important to set “Custom” in Timeline Resolution and put the number of pixels of the two dimensions in the boxes below as observable in the figure before clicking on “Create”.

After this, can be directly opened Fusion (the red selection in the figure 3.18), which is the section of the program dedicated to visual effects (VFX). Here, the most important panels are the node compositor (blue) to elaborate our videos, 2 view windows (green) to see the effects of a specific node on the image, and the Inspector (yellow), which allows checking and edits the attributes given to the effect of the selected node.

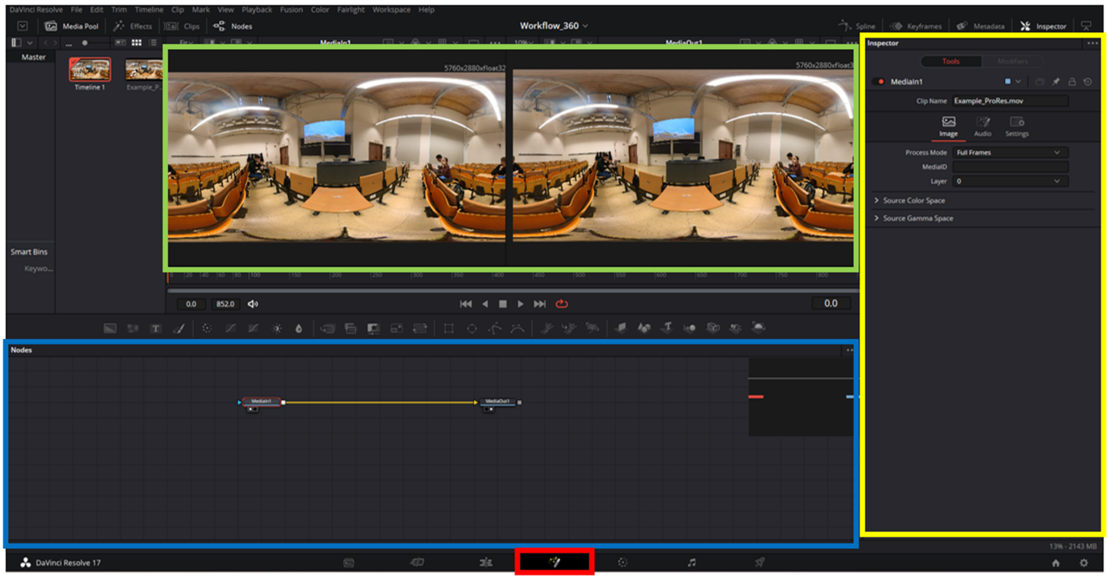
Now we’ll see the final disposition of the nodes to correct this image, an will be described the role of every node.



- MediaIn

  This node represents the input video file as it was before VFX (in this case the equirectangular).

- PanoMap



MediaIn1

Clip Name Example_ProRes.mov

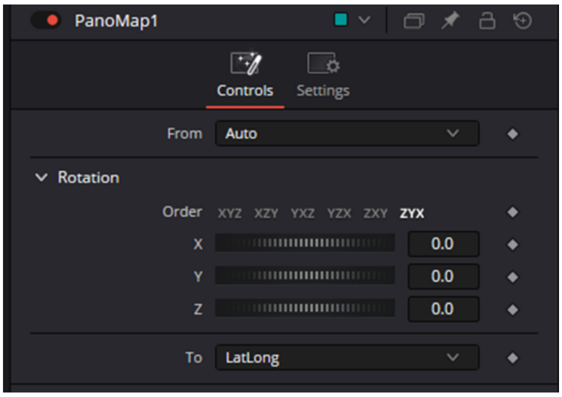
Process Mode Full Frames

MediaID

Layer 0

Figure 3.18: The Fusion section of DaVinci Resolve.

A VR-type node that allows reorienting the footage with a specific angle in the three directions of the space. The most frequent use of it is to do a pan, so modify the Y value to change the initial frontal direction. it's not mandatory using it, and it can be placed also at the end of the graph.



From Auto

Rotation

Order XYZ XZY YXZ YZX ZXY **ZYX**

X 0.0

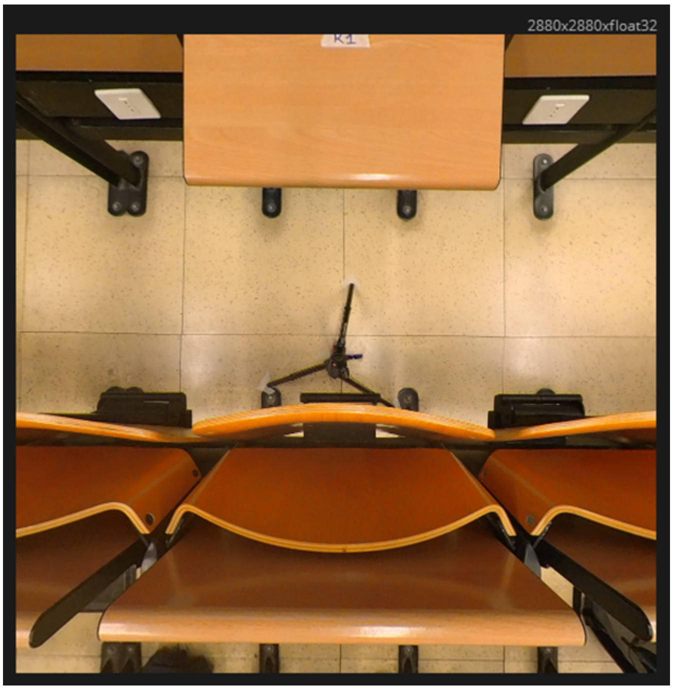
Y 0.0

Z 0.0

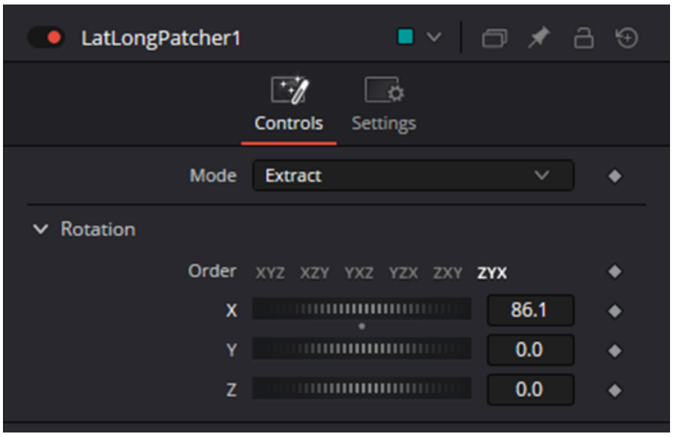
To LatLong

- LatLongPatcher1

A VR-type node with the function to extract (in the Extract Mode selected in the inspector), from an equirectangular 360 video, a square portion of the surface of the viewing sphere.



For the purpose of removing the tripod, must be taken the portion of the video's sphere with the tripod. In this node, the inspector is similar to the one node PanoMap, because also here must be selected an angle on the three axes where center the selected zone.



LatLongPatcher1

Mode Extract

Rotation

Order XYZ XZY YXZ YZX ZXY **ZYX**

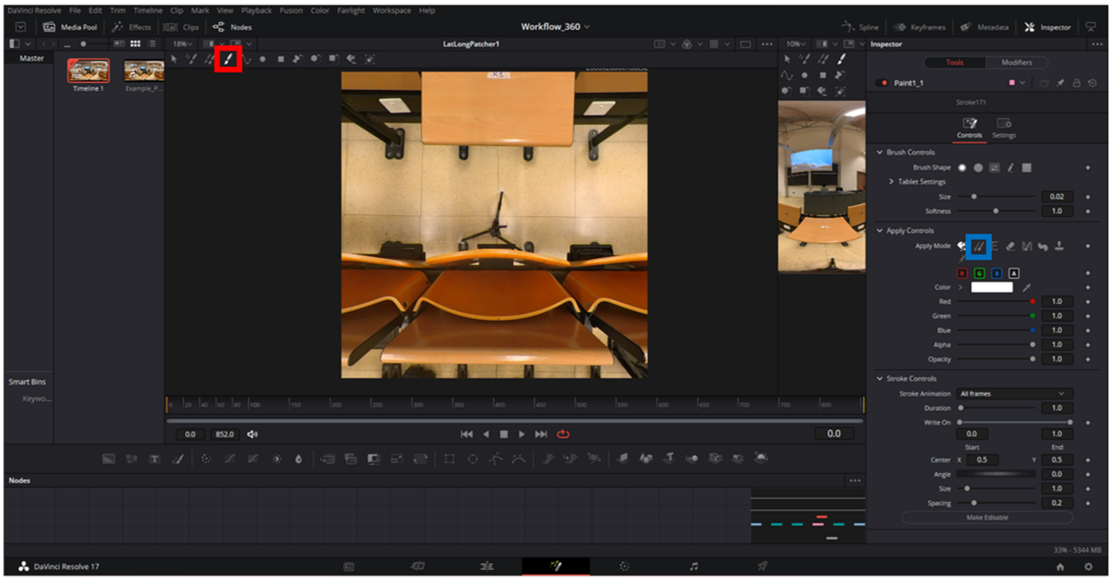
X 86.1

Y 0.0

Z 0.0

This node has been identified with also its number because it will also be used again later with a different function.

- Paint



Paint1_1

Stroke171

Brush Controls

Brush Shape

Tablet Settings

Size 0.02

Softness 1.0

Apply Controls

Apply Mode

Color

Red 1.0

Green 1.0

Blue 1.0

Alpha 1.0

Opacity 1.0

Stroke Controls

Stroke Animation All frames

Duration 1.0

Write On Start 0.0 End 1.0

Center X 0.5 Y 0.5

Angle 0.0

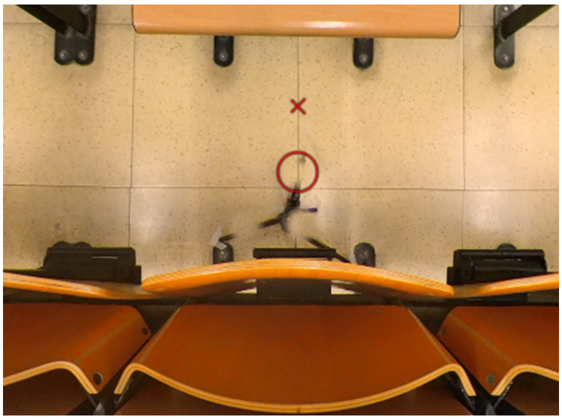
Size 1.0

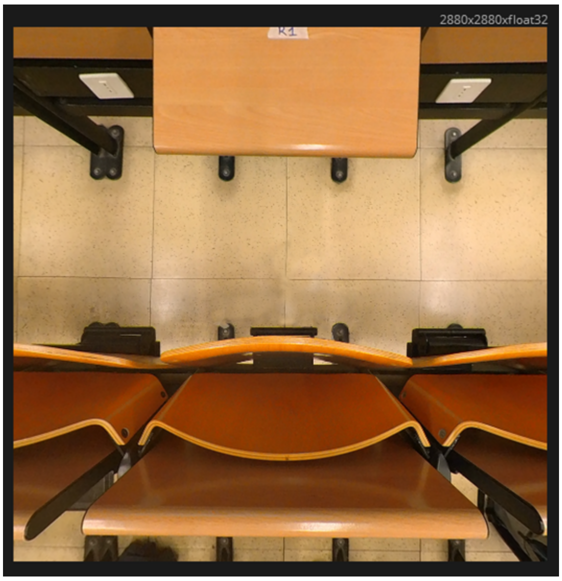
Spacing 0.2

Figure 3.19: The settings of Paint node.

The effective elimination of the tripod is done in this node, cloning parts of the floor around the tripod and overlapping them at the image parts where the tripod appears.

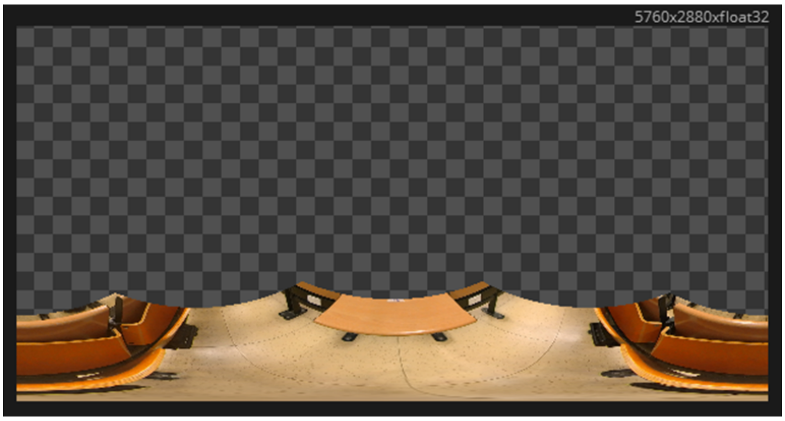
Must be selected Stroke mode in the selection bar on the top of the viewing window (3.19) then, in the inspector, select "Clone" as Apply Mode (blue in the figure). It is important the selection "All frames" in Stroke Animation, fundamental to maintain the effect during the whole video. Now the cloning is done by doing Alt+click in the reference point and, after, clicking on the point where to paste the circle selection. It must be done a few times, also trying to change the dimensions and the softness of the stroke until the result will be considered satisfying.





- LatLongPatcher2

  At this point, the portion of the sphere clipped with the first node of the LatLongPatcher must be taken back into equirectangular. It can be done by putting in series a second node of the same type with the same angles as the first but with the Mode set to "Apply". The output of this node, however, will not be the entire equirectangular image but only this previously cropped portion.



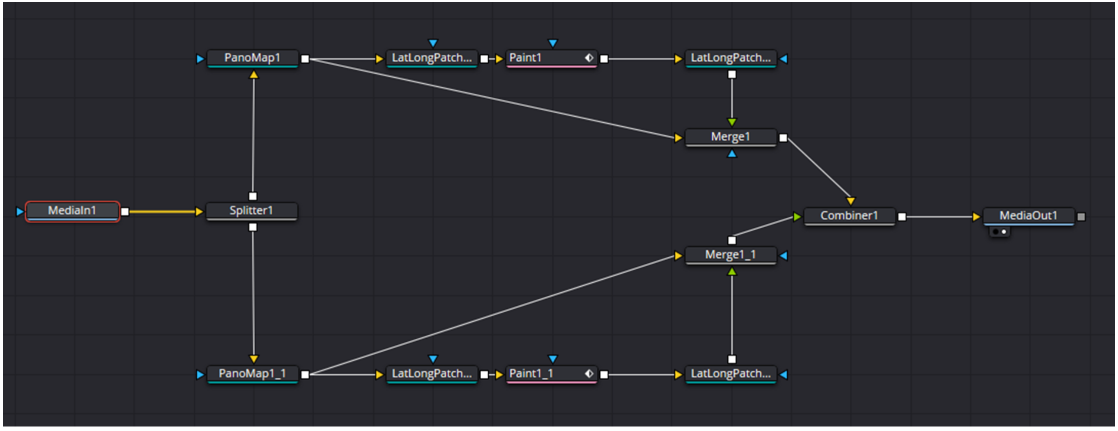
- Merge

  With this node, will be reunited the post-produced part of the image with the rest of the original one. It accepts three inputs: Foreground, Background, and Effect Mask. In this case, the last one should not be used. The original image will be used as the background, and the foreground will be the output of the second LatLongPatcher.

- MediaOut

  The final node of the output image, which corresponds to the output of the previous Merge node.

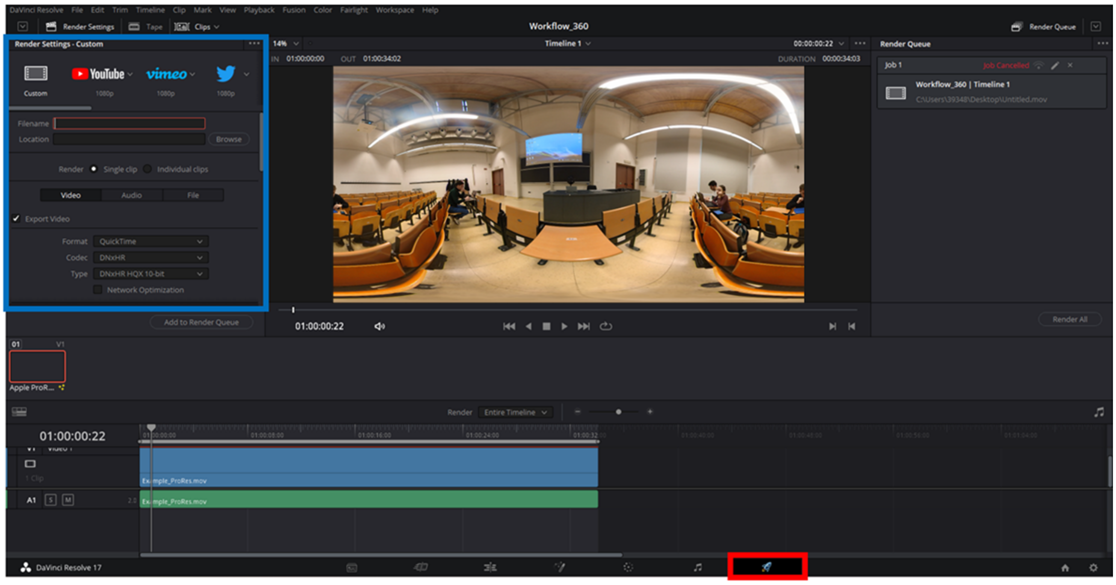
In the case of 3D footage, the process is similar. In the next picture, can be seen how to set the fusion node editor with a 3D image.



The process is the same, but separating the two images of the different eyes with the Stereo type node "Splitter" with the Split property set to "Vertical". It returns the two separated images, that will be processed in parallel. After the cancellation in both, they will be reunited with the Stereo type node "Combiner" with the

Combine property set to "Vertical". It is important to set the two PanoMap nodes with the same angles and, in the paint phase, recreate the floor in a way as similar as possible between the two images. Little differences between them will cause a perception of visual dirtiness during the reproduction in the HMD.

After the cancellation process on fusion can be done the exporting.



Filename
Location
Render ● Single clip ○ Individual clips
☑ Export Video
Format QuickTime
Codec DNxHR
Type DNxHR HQX 10-bit
□ Network Optimization

IN 01:00:00:00 OUT 01:00:34:02
00:00:00:22 DURATION 00:00:34:03
01:00:00:22

Render Queue
Job 1 Job Cancelled
Workflow_360 | Timeline 1
C:\Users\39348\Desktop\Untitled.mov

Render Entire Timeline
01:00:00:22
Example_ProRes.mov
Example_ProRes.mov

Figure 3.20: The Deliver section of DaVinci Resolve.

In the "Deliver" section (red in the figure 3.20) must be selected a codec similar to Apple ProRes to export. The reason for this choice is that for computationally complex operations such as the paint on Fusion that we used to remove the tripod, exporting directly to a highly compressed codec (e.g. H.264 or H.265) could cause excessive slowness or a program crash during the export. Must be therefore decided to a moderately compressed codec. Apple ProRes is fine, but it is only available with Apple devices (with a Windows machine like the one used here we can only play it but not encode it). Two good choices are the following:

- Format “Quick Time”, Codec “DNxHR”, Type “DNxHR HQX 10-bit”;
- Format “Quick Time”, Codec “GoPro CineForm”, Type “YUV 10-bit”.

A bit depth of 10 was chosen for both because it corresponds to that of the stitched ProRes file generated by the Insta360 software. The choice of 12 would create unnecessarily larger files. In the Render Settings (blue in the figure), in addition to the codec, can be set the filename and the path to which to export the file. Will be verified that the resolution is set to “Custom” and contains the dimension values in pixels equivalent to those of our equirectangular footage. It will be necessary to verify that also the framerate corresponds to the same as the original file. In "Advantage settings", to have the best possible quality, will be set the item "Data Levels" to "Full" and select "Force sizing to the highest quality". In this case, it can be disabled the audio export to save computational speed.

The last two clicks to do are on "Add to Render Queue" and then on "Render All" under the render queue to start rendering.

When the rendering is finished, there are still two important steps to perform before having the file ready to play in the viewer: encoding in a lossy codec and injecting metadata for 360 video.

The first step is to be performed with FFmpeg, an open-source framework that allows to do video file conversions from the command line of a computer [31].

The videos should preferably be encoded in H.265 because they would take up less size with the same quality output compared to H.264.

The maximum resolutions officially supported by Oculus Quest 2, which is the viewer that we have available, are as follows [32]:

Codec	Resolution	Fps
H.264	6200x3100 4300x4300	60
H.264	8192x4096 5760x5760	30
H.265	8192x4096 5760x5760	60

Apparently, just one type of shot done with the used cams is not compatible with what is reported, and they are the 6K 3D shots of the Insta360 Pro. Trying to upload them on the Oculus, however, for a reason for which hasn't been found an explanation even with research, the viewer can reproduce them if they are encoded in H.264. For this reason, will be encoded with FFmpeg all the 2D footage to H.265 and the 3D footage to H.264. Through an online tutorial [33] were found already made FFmpeg codes, which allow encoding any video in mp4 H.264 or H.265 keeping the video framerate and resolution intact and maintaining the other technical parameters in agreement with the quality recommended on the official Oculus website [32].

For H.264:

```
ffmpeg -i "input.mov" -c:v libx264 -preset slow -crf 18 -maxrate 100M -bufsize
200M -pix_fmt yuv420p -an -movflags faststart "output.mp4"
```

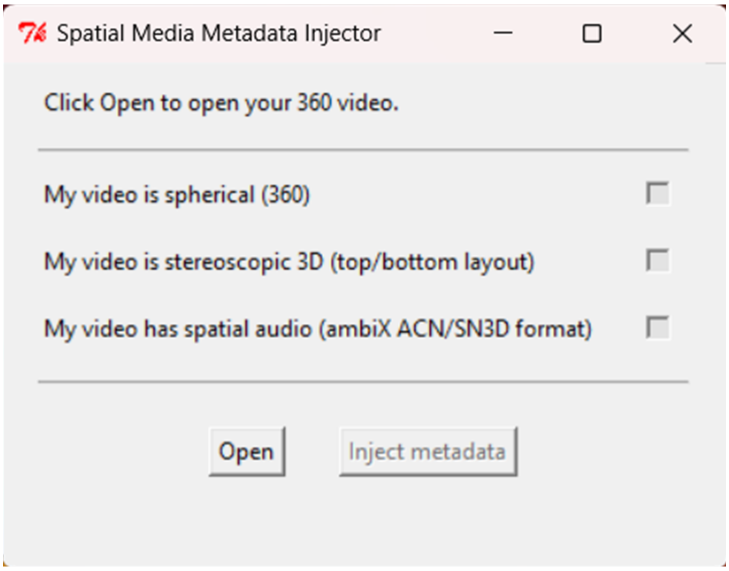
For H.265:

```
ffmpeg -i "input.mov" -c:v libx265 -preset slow -crf 18 -maxrate 100M -bufsize
200M -pix_fmt yuv420p -an -movflags faststart "output.mp4"
```

Obviously, *"input.mov"* must be replaced with the path of the file to encode.

One last precaution can be made regarding hardware acceleration. These codes work on all PCs because they encode video using only the CPU. They can be enhanced with different commands to use GPU encoding. In this case, having machines with Nvidia GPUs, the use of hardware acceleration would considerably accelerate export processes. To do it we have to replace *"libx264"* with *"h264_nvenc"* and *"libx265"* with *"h265_nvenc"* [34]. Unfortunately, the Nvidia hardware encoding of H.264 supports a maximum resolution of 4096x4096. Then, in the case of the 3D 6K videos, they must be encoded in H.264 using the normal command with *"libx264"*. Here, can effectively be used the Nvidia hardware encoding in H.265 for our 2D clips.

Once the file has been encoded, the last thing left to do is inject metadata into it to make it recognizable as a 360 video by any platform, including the Oculus Quest 2 video player app. To do this must be used the free program “Spatial Media Metadata Injector”.



Spatial Media Metadata Injector

Click Open to open your 360 video.

My video is spherical (360) □

My video is stereoscopic 3D (top/bottom layout) □

My video has spatial audio (ambiX ACN/SN3D format) □

The use is very simple: after importing the file with the "Open" button, must be selected "My video is spherical (360)" and, in the case of 3D video, also select the second option. The third option, relating to spatial audio, should not be selected because, in this first phase of the current project, audio will be played externally from the video file. Then, by clicking on “Inject metadata”, the final file is saved.

Chapter 4

Applications: scene reproduction inside the Audio Space Lab

The material produced in the chapter 3 is subsequently used to reconstruct the ecological scenes in which to perform listening tests in the Audio Space Lab. After an overview on the equipment present in the laboratory, in this chapter, the mathematical steps for obtaining the sound scene tracks are described.

4.1 Audio Space Lab Structure & Reproduction System

The Audio Space Lab (figure 4.1) a laboratory in the Energy Department of the Polytechnic of Turin. This is a room with a reverberation time of 0.17 s at mid-frequencies. Inside it, the present installation is composed of 18 loudspeakers. The main array of loudspeakers is spherical and consists of 16 Genelec 8030B units arranged in three rings: one of eight units at the height of the listener's ears and two of 4 units placed above and below. The loudspeakers in the central ring are equally spaced along the equatorial circumference of the sphere, therefore they have an azimuth difference of 45° starting from the first one which is placed in the frontal direction of the user. The other two rings are arranged to have the elevation angles of the loudspeakers at 45° and -45° respectively, and with the azimuths of 45°, 135°, 225° and 315° for both rings. The last two units, of the Genelec 8351B, are placed frontally in two corners of the room and have the function of subwoofers. The characteristics of the two loudspeaker models used, taken from the official Genelec datasheets [35] [36], are compared in the table 4.1.

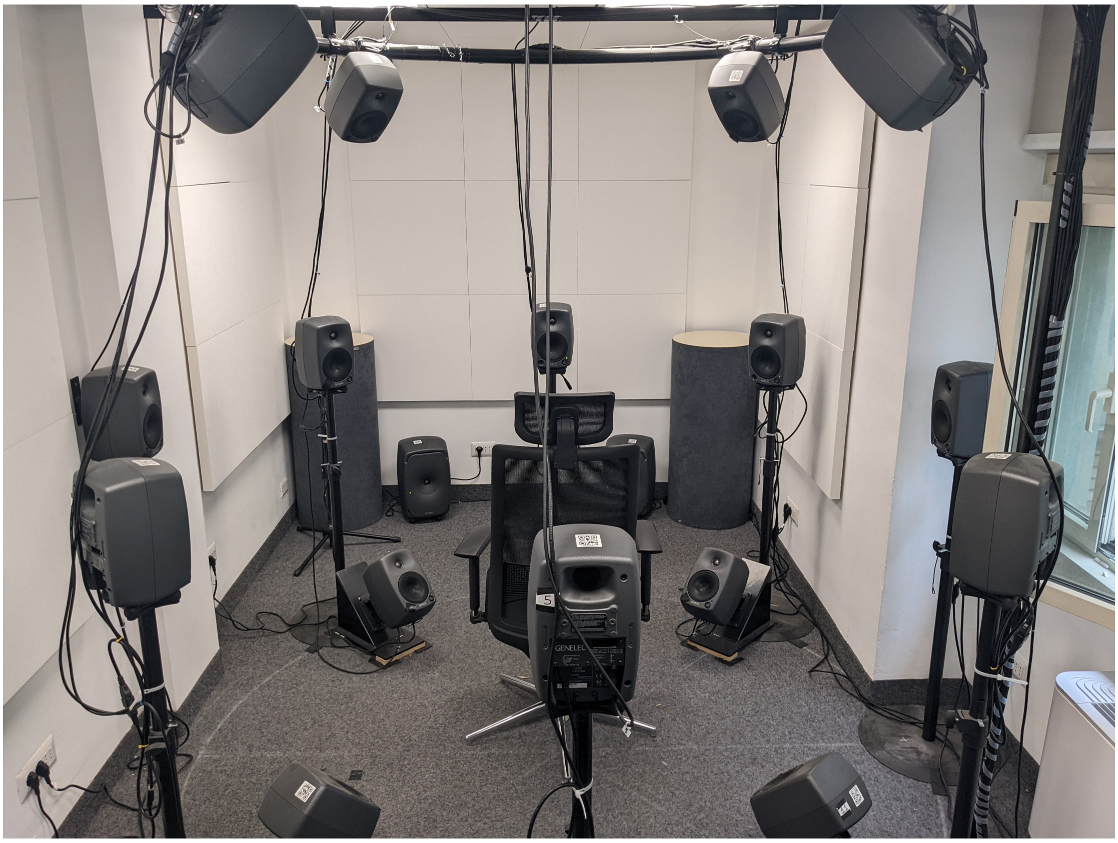


Figure 4.1: The structure of the Audio Space Lab at Politecnico di Torino.

The system of loudspeakers is driven by the audio card Antelope Orion 32 4.2. This device allows the reproduction of the audio played on a DAW on an array of loudspeakers with a maximum of 32 channels simultaneously.

Figure 4.2: The Antelope Orion 32 audio card.

The 3rd-order Ambisonics tracks are played by the blocks-structure DAW Bidule. In the figure 4.3 there is an approximate scheme with the main blocks that were set in the program. The most important of these, which is the one that decodes to adapt the Ambisonics B-format to the loudspeakers array used, is the “AIIRADecoder”. This block contains a transformation matrix which is calculated starting from the SN3D seen in the table 2.2 and from the positions of the various loudspeakers in

	Genelec 8030B	Genelec 8351B
Dimensions	Height with stand: 299mm Stand's height:14mm Width: 189mm Depth: 178mm	Height with stand: 454mm Stand's height: 21mm Width: 387mm Depth: 278mm
Weight	5.6 Kg	14.3 Kg
Lower cut-off frequency	$\leq$ 55 Hz (-6dB)	$\leq$ 32 Hz (-3dB)
Upper cut-off frequency	$\geq$ 21 kHz (-6dB)	$\geq$ 43 kHz (-3dB)
Accuracy of frequency response	58 hZ - 20 kHz (±2 dB)	38 Hz – 20 kHz (±1.5 dB)
Maximum short term sine wave acoustic output on axis in half space, averaged from 100 Hz to 3 kHz at 1 m	100 dB SPL	113 dB SPL
Self generated noise level in free space (1 m on axis)	$\leq$ 5 dBA	$\leq$ 10 dBA
Drivers	Bass: 130 mm cone Treable: 19 mm metal dome	Bass: dual 218x108 mm obround cones Midrange: 130 mm cone Treable: 25 mm metal dome

Table 4.1: Comparing properties of Genelec 8030B and Genelec 8351B

the listening array set by the user.

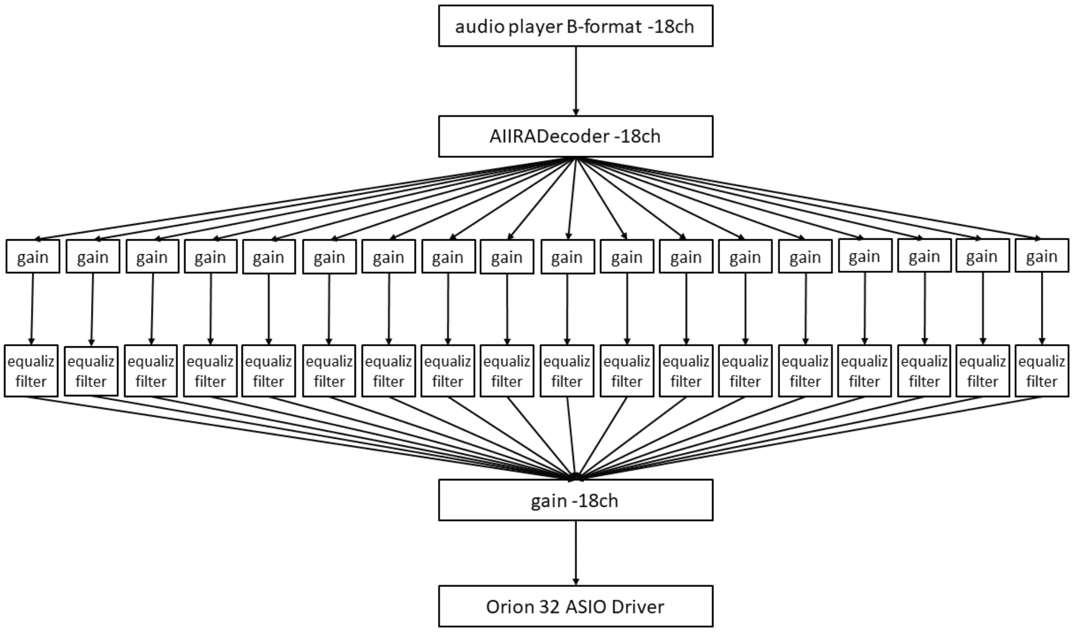


Figure 4.3: Settings of Bidule used to play 3rd order Ambisonics on the ASL loudspeakers-array.

4.2 Creation of a scene for speech-in-noise tests

The ecological sound scenes to be reproduced in the ASL are created using a MATLAB routine. The starting signals are the target and the noise recorded in the anechoic version, the inverse sweep, and the two recorded sweeps relating to the target and noise sources. The main steps to be performed are the following:

1. generate target and noise RIRs from the sweeps with the procedure mentioned in paragraph 2.1.2;
2. convolve the target test anechoic signal with its relative RIR;
3. convolve the noise test signal anechoic with its relative RIR;
4. sum the two tracks obtained in points 2 and 3 while adding a proper gain to the auralized target to insert the desired SNR.

Chapter 5

Conclusions and future perspective

Currently, the benefits of HAs for HIOAs are often limited, mainly due to current audiometric testing standards, which do not take into account many aspects of real-life hearing. For this reason, new audiometric testing standards are being developed which involve the audiovisual reproduction in VR of FAEs, like the ones that will be developed in the ASL at the Polytechnic of Turin.

Two literature searches were done: the first concerning the environments most considered in audiometry to identify the FAEs to be reproduced, and the second to search for spatial audiovisual scenes databases similar to the one that will be created for the current project. After having confirmed that no existing database fully satisfies the needs of this study, the environments in which to do the measurements to create a new database were identified. In each environment, two listener positions and set of noise positions were defined. The positions were decided according to the criteria of perceived acoustic complexity and the azimuth angles with maximum and minimum spatial release from masking.

At the current state of the works, all the necessary measurements were done in the first two identified environments. All the 3OA RIRs relative to the noise and target positions were measured with respect to those of the listeners and the 360 videos were filmed in 2D and 3D for all the configurations taken into account. The acoustical parameters that characterize these two environments were also calculated from the result of RIRs measured in random positions inside the environment. Finally, the videos were post-produced to mask the presence of the tripod under the camera.

The created 26 audiovisual scenes will be used to perform the first ecological listening tests of speech-in-noise in the ASL. The first tests will not be performed on HIOAs, which is the main target of the project, but on normal-hearing people. These first tests will be aimed at validating the environments, therefore at understanding whether there are differences in results of perceptual tests done with and without visual content. If they will be successful, the content production workflow described here will be considered also for subsequent environments. An anechoic chamber, to implement the reference condition, and a café have already been confirmed as the next environments in which to perform the measurements.

A limit that has not been considered in this phase of the work, but which will be a priority in the studies of the next ones, is audio and video synchrony. The current state of the installation provides for an independent reproduction of audio and video. The video will first be loaded into the HMD Meta Quest 2 and played through the Oculus TV app. At the same time, the audio signal will be emitted by the ASL loudspeakers using the system described in the chapter 4. At the moment, the audio content of the captured AV scenes has no elements that need to be synchronized with the visual content as, for example, no lip movements are visible (all audible speech is visually represented by the TalkBox, the dodecahedron, the HaTS, and people acting to talk hiding their lips).

Regarding this problem, the short-term objectives defined are two:

- be able to synchronize audio and video reproduction in ASL;
- make synchronized audio and video recordings.

As regards the first objective, a system has already been created, but still needs to be validated. It will allow, through Unreal Engine, to synchronize the timecode of the execution of the audio and video files in the two different platforms used, to be able to reproduce these contents in a synchronized way.

The second objective is also in the process of being achieved. Since the Insta360 Pro camera was provisionally lent for filming these first two environments, the subsequent work will be done exclusively with the Insta360 ONE X2. Indeed, by using 3D printing, for the latter mentioned camera, a special support (figure 5.1) has already been created, which will allow the audio and video filming centers to be kept on the same vertical axis in the same location at the same time. In particular, it was decided to place the Zylia above the video camera to minimize the distance between the audio and video recording centers. The use of this equipment will allow to register in real-time and at the same time audio and video, keeping a precise match, so allowing for greater realism and a greater feeling of immersion in the environment.

It will also cause a greater complexity of the work compared to what has been done up to now, as it will be further necessary:

- to produce an instantaneous peak signal to synchronize audio and video, like a hand clap or a clapperboard;

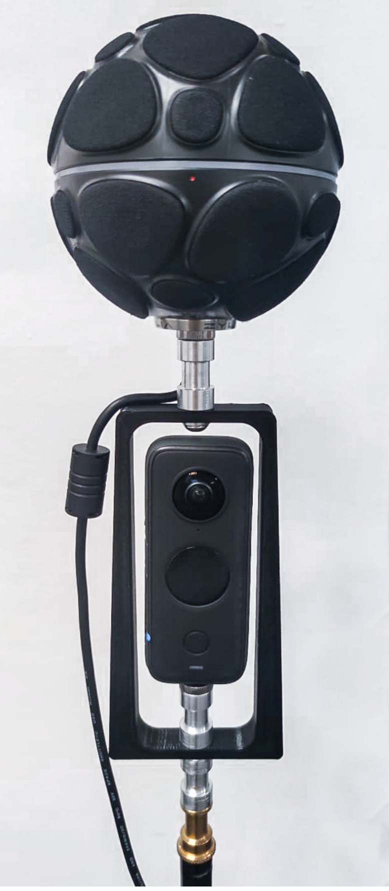


Figure 5.1: The stand that will be used to make synchronized audiovisual recordings in the environments that will be measured in the future.

- to eliminate, with the same process seen in the section 3.5, also the PC that will be used to record with the Zylia;
- to eliminate the Zylia, placed above the video camera, with a different post-production process from the one used to eliminate the tripod;
- to compute filters to be applied on the recorded audio signals to compensate for the influence of the camera on the sampled sound field.

A further development idea to improve the quality of the work concerns the scenography placed under the video camera. As already seen, in the shootings taken in the first two environments the tripod was removed from the floor to have greater immersion in the scene and show as little as possible the elements used for the recordings. However, the advantages of this removal are limited, since a very unreal element remains: the visual absence of the chair placed under the user who, if he looks down, will perceive himself as sitting suspended in the air. This problem, which was not initially taken into consideration, can be bypassed in the measurements of the subsequent environments through scenographic tricks. A simple but effective one could be the creation of a hole in a chair identical to those used in the environment, through which the tripod rod that holds the video camera would be passed. In the post-production phase, the operations to be done would become much simpler, because it would be sufficient to mask that single hole by cloning a piece of the chair's surface. Not only the content of the video would be more consistent with reality, but there would also be a much lower risk of creating defects in the image due to the removal of the tripod on a surface that is not too homogeneous such as the tiled floors in the classroom and in the conference hall. The result would be more stable even in 3D videos because the risk of imperfect perception caused by little differences in the two images would be minimized.

Bibliography

[1] *Deafness and hearing loss — who.int.* `https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss`. [Accessed 20-Feb-2023] (cit. on p. 1).

[2] *World report on hearing — who.int.* `https://www.who.int/publications/i/item/9789240020481`. [Accessed 20-Feb-2023] (cit. on p. 1).

[3] Abby McCormack and Heather Fortnum. «Why do people fitted with hearing aids not wear them?» In: *International journal of audiology* 52.5 (2013), pp. 360–368 (cit. on p. 1).

[4] Richard H Wilson and Wendy B Cates. «A comparison of two word-recognition tasks in multitalker babble: Speech Recognition in Noise Test (SPRINT) and Words-in-Noise Test (WIN)». In: *Journal of the American Academy of Audiology* 19.07 (2008), pp. 548–556 (cit. on p. 1).

[5] Jens Cubick and Torsten Dau. «Validation of a virtual sound environment system for testing hearing aids». In: *Acta Acustica united with Acustica* 102.3 (2016), pp. 547–557 (cit. on p. 2).

[6] *ISO 3382-2:2008 — iso.org.* `https://www.iso.org/standard/36201.html`. [Accessed 09-Mar-2023] (cit. on pp. 4, 6).

[7] *ISO 3382-1:2009 — iso.org.* `https://www.iso.org/standard/40979.html`. [Accessed 09-Mar-2023] (cit. on pp. 4, 6).

[8] *Impulse Response 2013; Studio Six Digital — studiosixdigital.com.* `https://www.studiosixdigital.com/audiotools-modules-2/acoustic-analysis-modules/impulse_response/`. [Accessed 25-Feb-2023] (cit. on p. 8).

[9] Giso Grimm, Birger Kollmeier, and Volker Hohmann. «Spatial acoustic scenarios in multichannel loudspeaker systems for hearing aid evaluation». In: *Journal of the American Academy of Audiology* 27.07 (2016), pp. 557–566 (cit. on pp. 8, 9, 26, 35).

[10] Maartje ME Hendrikse, Gerard Llorach, Volker Hohmann, and Giso Grimm. «Movement and gaze behavior in virtual audiovisual listening environments resembling everyday life». In: *Trends in Hearing* 23 (2019), p. 2331216519872362 (cit. on pp. 8, 26, 35).

[11] Giso Grimm, Joanna Luberadzka, and Volker Hohmann. «Virtual acoustic environments for comprehensive evaluation of model-based hearing devices». In: *International journal of audiology* 57.sup3 (2018), S112–S117 (cit. on pp. 9, 26).

[12] Ayham Zedan, Tim Jürgens, Ben Williges, Birger Kollmeier, Konstantin Wiebe, Julio Galindo, and Thomas Wesarg. «Speech intelligibility and spatial release from masking improvements using spatial noise reduction algorithms in bimodal cochlear implant users». In: *Trends in Hearing* 25 (2021), p. 23312165211005931 (cit. on pp. 9, 26).

[13] Ayham Zedan, Tim Jürgens, Ben Williges, David Hülsmeier, and Birger Kollmeier. «Modelling speech reception thresholds and their improvements due to spatial noise reduction algorithms in bimodal cochlear implant users». In: *Hearing Research* 420 (2022), p. 108507 (cit. on pp. 9, 26).

[14] Annelies Devesse, Astrid van Wieringen, and Jan Wouters. «AVATAR assesses speech understanding and multitask costs in ecologically relevant listening situations». In: *Ear and Hearing* 41.3 (2020), pp. 521–531 (cit. on pp. 10, 26).

[15] Chris Oreinos and Jörg M Buchholz. «Evaluation of loudspeaker-based virtual sound environments for testing directional hearing aids». In: *Journal of the American Academy of Audiology* 27.07 (2016), pp. 541–556 (cit. on pp. 10, 26).

[16] Douglas S Brungart, Benjamin M Sheffield, and Lina R Kubli. «Development of a test battery for evaluating speech perception in complex listening environments». In: *The Journal of the Acoustical Society of America* 136.2 (2014), pp. 777–790 (cit. on pp. 10, 26).

[17] Gerard Llorach, Frederike Kirschner, Giso Grimm, Melanie A Zokoll, Kirsten C Wagener, and Volker Hohmann. «Development and evaluation of video recordings for the OLSA matrix sentence test». In: *International Journal of Audiology* 61.4 (2022), pp. 311–321 (cit. on p. 11).

[18] Lubos Hladek and Bernhard U Seeber. «Audiovisual models for virtual reality: Underground station». In: *Fortschritte der Akustik–DAGA'22*. 2022 (cit. on pp. 11, 19).

[19] Adam Weisser, Jörg M Buchholz, Chris Oreinos, Javier Badajoz-Davila, James Galloway, Timothy Beechey, and Gitte Keidser. «The ambisonic recordings of typical environments (ARTE) database». In: *Acta Acustica United With Acustica* 105.4 (2019), pp. 695–713 (cit. on pp. 12, 26).

[20] Georg Götz, Sebastian J Schlecht, and Ville Pulkki. «A dataset of higher-order Ambisonic room impulse responses and 3D models measured in a room with varying furniture». In: *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE. 2021, pp. 1–8 (cit. on p. 15).

[21] Marc Ciufo Green and Damian Murphy. «EigenScape: A database of spatial acoustic scene recordings». In: *Applied Sciences* 7.11 (2017), p. 1204 (cit. on pp. 18, 26).

[22] Steven Van De Par et al. «Auditory-visual scenes for hearing research». In: *Acta Acustica* 6 (2022), p. 55 (cit. on pp. 19, 26).

[23] Adelbert W Bronkhorst. «The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions». In: *Acta Acustica united with Acustica* 86.1 (2000), pp. 117–128 (cit. on p. 21).

[24] Giuseppina Emma Puglisi, Anna Warzybok, Arianna Astolfi, and Birger Kollmeier. «Effect of reverberation and noise type on speech intelligibility in real complex acoustic scenarios». In: *Building and Environment* 204 (2021), p. 108137 (cit. on pp. 22–24, 26, 35).

[25] Daniel Arteaga. «Introduction to ambisonics». In: *Escola Superior Politècnica Universitat Pompeu Fabra, Barcelona, Spain* (2015), p. 21 (cit. on pp. 25, 28–30).

[26] Angelo Farina. «NUOVI METODI E STRUMENTI DI MISURA PER LO STUDIO DELL'ACUSTICA DEI TEATRI STORICI ITALIANI». In: () (cit. on pp. 27, 28).

[27] Francesca Ortolani. «Introduction to Ambisonics». In: *Ironbridge Electronics* (2015) (cit. on pp. 28, 29, 32).

[28] *File:Spherical Harmonics deg5.png - Wikimedia Commons — commons.wikimedia.org*. `https://commons.wikimedia.org/wiki/File:Spherical_Harmonics_deg5.png`. [Accessed 15-Feb-2023] (cit. on p. 32).

[29] Adam Weisser, Jörg M Buchholz, and Gitte Keidser. «Complex acoustic environments: Review, framework, and subjective model». In: *Trends in hearing* 23 (2019), p. 2331216519881346 (cit. on p. 35).

[30] *3.3.1 [Beginner] Video stitching by Stitcher - Pro User Manual*. `https://onlinemanual.insta360.com/pro1/en-us/video/postproduction/1`. (Accessed on 01/30/2023) (cit. on p. 54).

[31] *FFmpeg — ffmpeg.org*. `https://ffmpeg.org/`. [Accessed 06-Feb-2023] (cit. on p. 68).

[32] *Encoding immersive videos for Meta Quest 2 | Meta Quest for Creators — creator.oculus.com.* `https://creator.oculus.com/getting-started/media-production-specifications-for-delivery-to-meta-quest-2-headsets/?locale=it_IT`. [Accessed 06-Feb-2023] (cit. on pp. 68, 69).

[33] *Oculus Quest 2: What's the MAX 360° Video Resolution? Render Settings Tutorial & my Quest 2 Review — youtube.com.* `https://www.youtube.com/watch?v=a4Wm2yX1q2o&t=4s`. [Accessed 06-Feb-2023] (cit. on p. 68).

[34] *Using FFmpeg with NVIDIA GPU Hardware Acceleration :: NVIDIA Video Codec SDK Documentation — docs.nvidia.com.* `https://docs.nvidia.com/video-technologies/video-codec-sdk/ffmpeg-with-nvidia-gpu/`. [Accessed 06-Feb-2023] (cit. on p. 69).

[35] *8030B - Genelec.com — genelec.com.* `https://www.genelec.com/previous-models/8030b`. [Accessed 21-Feb-2023] (cit. on p. 71).

[36] *8351B - Genelec.com — genelec.com.* `https://www.genelec.com/8351b`. [Accessed 21-Feb-2023] (cit. on p. 71).