POLITECNICO DI TORINO

Master of Science
Cinema and Media Engineering

Master's Thesis

Improvement of speech clarity on flat screen TVs for elderly people with presbycusis without otological pathologies and for young non-native speakers

Supervisors
Prof.ssa Arianna Astolfi[1]
Prof. Marco Masoero[1]
Prof.ssa Louena Shtrepi[1]
Dott. Fabrizio Riente[2]
Dott. Andrea Albera[3]
[1]: DENERG, [2]: DET, [3]:UniTO

Candidate
Lorenzo Meli

April 2021

Contents

List of Figures

List of Tables

Introduction

In recent years, television (TV) manufacturers have developed larger and thinner screens. The limited space available in a flat TV affects the placement of loudspeakers. Moreover, manufacturers have miniaturized and installed loudspeakers facing down or towards the wall, diametrically opposite to spectators. These facts lead to a worsening of the audio quality listened to by the viewer. Therefore, speech intelligibility is degraded along with transmission problems that are already present in many television programs. One solution could be to integrate the home environment with a hardware component, such as a soundbar. Nevertheless, most end-users rely only on speakers embedded in television, both for economic reasons and poor audio culture.

This research project funded by RAI (Radiotelevisione Italiana) aims to create a system capable of improving the audio quality heard by the television broadcaster customers on the various platforms (DTT, SAT, and IP) in cases where the television receiver is flat. In particular, the research aims to develop a transfer function (t.f.) to be implemented on a Digital Audio Optimizer (DAO), which is characterized in terms of Digital Signal Processor (DSP) to compensate for degradation factors that impair the speech intelligibility of the audio message.

As part of the previous project "Analysis and improvement of the sound quality perceived by a flat-screen TV," a transfer function called "Heavy" was proposed. It was validated by the Politecnico di Torino and implemented in the DSP Junger D*AP4 FLX[1], which has improved the listening experience about the sound quality perceived on most of the flat TV screens on the market. More precisely, the t.f. proposal improves RAI products' signal clarity, divided into three genres: Spoken, Sport, and Music. In particular, the analysis of the results of a subjective listening test based on the Subjective Difference Grade (SDG), which expresses in percentage the improvement or deterioration of the track treated with the t.f. compared to the original, it showed an improvement of 25.3% on average on three different television models and genres.

RAI, interested in further studying the positive aspects of the implemented function, intends to evaluate the benefits of t.f. "Heavy" and its modified version "Moderate." The research will focus on older adults with presbycusis and without otological pathologies and the case of young non-native speakers in collaboration with the otolaryngologists of the University of Turin. The previous investigations employed the model LG 55UK6100PLB, 55" (2 speakers, total power 20 W), which provided the best result considering commercial televisions. Thus, in this research, we will use the same device.

The TV will be placed in an anechoic room and in a non-anechoic environment (Audio Space Lab) of Politecnico di Torino, where we will perform subjective tests. For this purpose, the television's frequency response will be assessed in both rooms using the pulse response method. Furthermore, a reverberation time measurement and modal analysis of the Audio Space Lab will be conducted to validate it as a listening environment in conformity with ITU-R BS.1116-3[2] "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems." The same audio material previously used in other research projects will be employed during this investigation: 30 tracks extracted from Rai programs and diverse in three different genres. After selecting the audio material, the Junger D*AP4 FLX DSP will process the signal with the t.f. The dynamic equalizer and the limiter will be used among the available components not to exceed the loudness value established by the norm EBU-R128[3], equal to -23 $LUFS$. The transfer functions aim to amplify the frequencies useful to increase speech intelligibility by using dynamic equalization. In particular, the "Heavy" t.f. amplifies frequencies in the 1-5 kHz range and attenuates the spectrum extremes. On the other hand, the "Moderate" one does the same but in a more limited way. The investigation will involve two groups of people. The first one gathers elderly users, while young people who are not Italian native speakers constitute the second one. Both of them will be subjected to the subjective listening test following the method always indicated in the ITU-R BS.1116-3[2] and called "Double-blind triple-stimulus with hidden reference." The results will highlight that the objective initially set, equal to 20% of the improvement in speech clarity, is achieved for both groups of users tested. The present thesis will cover the following points:

The *First Chapter* describes the instrumentation used for all test phases, the hardware and software components, and the test's location. The Audio Space Lab is analyzed in all its features, emphasizing the aspects that led to subjective testing choices.

The *Second Chapter* covers the analysis of the TV frequency response used. Through the impulse response method, its characteristics are analyzed both in the anechoic chamber and in ASL. Then, the video material employed during the listening tests is described.

In the *Third Chapter*, there is an analysis of speech intelligibility based on scientific articles. The Digital Signal Processor setup and the two transfer functions implemented are described, with examples showing how they process the audio track in real-time.

The *Fourth Chapter* reviews the methodology used to test the two filters created. ITU-R BS.1116[2] is used as a guideline to create the best test conditions. The preliminary test consists of a liminar tonal audiometry, and the otolaryngologists perform this evaluation. The test analyzes the auditory participants' thresholds, allowing the doctors to understand if the test takers are suitable or not for the listening test. Furthermore, this chapter shows the interface the test subjects use to evaluate the processed audio track. Through the "Subjective Difference Grade," this evaluation is compared with the others and transformed into a percentage value.

The *Fifth Chapter* reports the results obtained from subjective tests through ANOVA statistical analysis and SDG values, divided by audio track and genre (Music, Speech, Sport).

The *Sixth Chapter* collects the final summary considerations based on the results of the previous chapter.

Part I

Experimental setup

Chapter 1

Equipment used and test location

The first chapter describes the hardware and software equipment used during the project. Figure 1.1 shows a general diagram of the instrumentation.

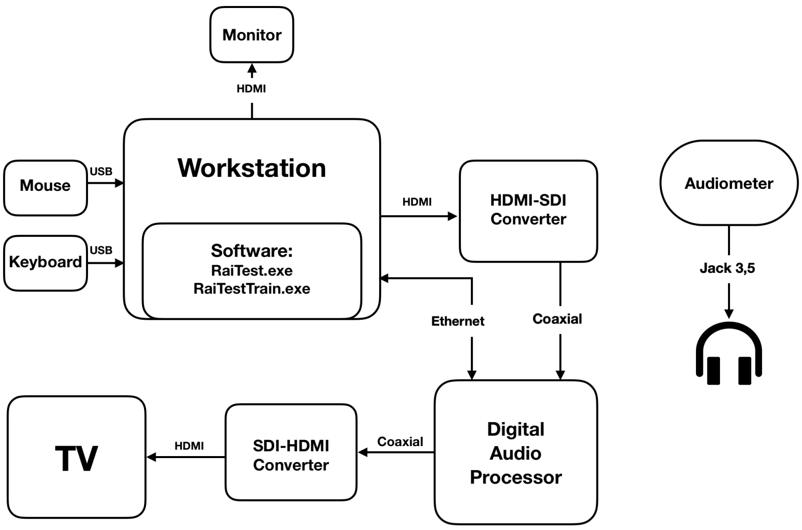


Figure 1.1: General instrumentation scheme.

1.1 Hardware

The following points show the hardware components employed during the reasearch activity.

1.1.1 Tv

The television used in this research is the LG 55UK6100PLB. It is a **55” Ultra HD Smart 4K Active HDR LED TV** with integrated Dolby Digital decoder and two speakers with a total power of 20 watts. It was chosen because more performing than the other devices used in the previous study [4].

Figure 1.2: LG TV LED 55” Ultra HD Smart TV 4K

1.1.2 Workstation

The workstation consists of a **3630 Dell Precision Tower** and a **28" Ultra HD 4K Dell S2817Q**. It was used to program the Audio Processor, to run test programs for users, and to send video files to the TV.

(a) *3630 Dell Precision Tower* (b) *28" Dell S2817Q*

Figure 1.3: Workstation used.

1.1.3 Digital Signal Processor

The TV station's program audio was processed through the **Junger D*AP4 FLX**[1] stereo digital audio processor. It is an electronic apparatus of the Digital Signal Processor (DSP) family, a hardware system that numerically processes digital signals after they have been sampled.

Figure 1.4: Junger D*AP4 FLX.

It performs operations such as equalization, compression, expansion, limitation, filtering of audio signal. Some of its characteristics are:

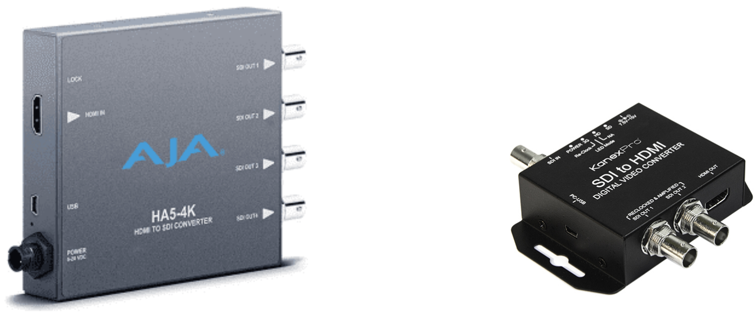
- Flexible feature set, configurable by licenses
- Dual stereo audio processing
- True peak limiter
- Optional Level Magic loudness management (optional) according to: ITU-R BS.1770 (all revisions), EBU R128, ATSC A/85, ARIB TR-B32, Free TV OP-59 and Portaria 354
- Loudness measurement supporting all worldwide standards
- On board interfaces: 2x AES/EBU I/O (XLR & BNC), Sync in, Sync out, 8 GPI/O
- Ethernet connectivity for set up and control via web browser
- External control via network or GPI/Os

For more information about the Digital Audio Processor, see Appendix B. The functionalities performed by the DSP in this study will then be explained.

1.1.4 Signal converters

The Junger DSP uses SDI connectors to connect to the workstation and TV, which have HDMI connectors. Two signal converters were used for this (Figure 1.5):

- **AJA HA5-4K**[5] for HDMI-to-SDI Workstation-Junger connection
- **Kanexpro SDI-SDHDXPRO**[6] for SDI-to-HDMI Junger-TV connection



(a) *AJA HA5-4K* (b) *Kanexpro SDI-SDHDXPRO*

Figure 1.5: Signal converters.

1.1.5 Audiometer

The otolaryngologist employs the audiometer **Inventis Triangle** for audiometric testing of TV users. It performs tonal audiometry in manual and automatic air-borne mode, in a frequency interval from 125 Hz to 8000 Hz. Depending on the needs, it is also possible to perform audiometry by bone. For more information about the audiometer, see Appendix A.

Figure 1.6: Junger D*AP4 FLX.

1.1.6 Various equipment

During the experiment, other equipment was used such as:

- Headphones supplied with the Audiometer

- 2x HDMI cables for connections between TV, Workstation and signal converters
- 2x BNC cables for connections between the DSP and the two signal converters
- 1x Ethernet cable between DSP and Workstation

1.2 Software

To submit tests to TV users, the softwares used are:

- RaiTest.exe
- RaiTestTrain.exe

Then, the results were analysed using::

- Matlab 2020
- Microsoft Excel

1.3 Location

1.3.1 Introduction

The study's protagonists are private users placed in conditions similar to those found in their living room at home. For this reason, the tests were not carried out in an entirely anechoic chamber. In this latter type of room, there are no reflections that interfere with those of the original signal. The user, or the microphone, receives the direct waves that come out from the speakers without conditioning due to objects or walls. Therefore, the room chosen is the **Audio Space Lab (ASL)** within the Department of Energy (DENERG) of the Polytechnic University of Turin. The room has good characteristics to perform this type of study, and the acoustics is not too distant from what a person would find in his living room. The sound space is influenced by the direct waves and those of the reverberating field[1] going to create stationary waves, which may modify the perception of the sound source based on the position of the source or of the listener.

[1]Sound waves that have received at least one reflection.

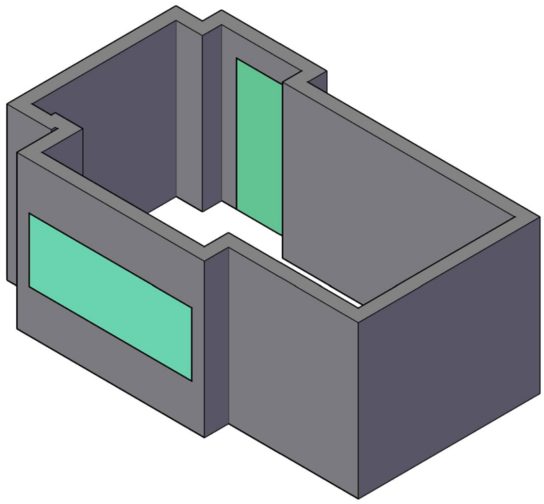


Figure 1.7: Audio Space Lab axonometry.

1.3.2 Stationary waves and acoustic nodes

A *stationary wave* is a wave whose amplitude is fixed in both time and space. It happens due to the interference between a direct wave and a reflected one. These waves propagate along the same line but in opposite directions. Moreover, in a stationary wave, the net transport of acoustic energy is zero. As Figure 1.8 shows,

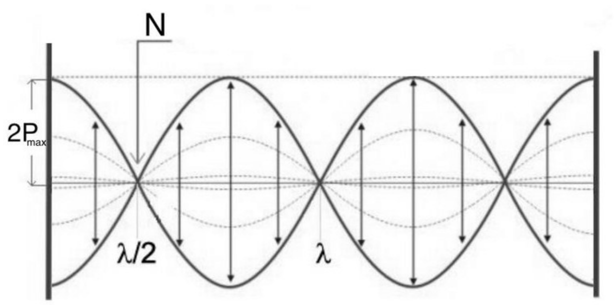


Figure 1.8: Example of stationary wave.

the interference is destructive in the points N (nodes) where deletion occurs, at points where x is an odd multiple of $\lambda/4$; instead, it is constructive at points A (antinodes or bellies) where the pressure is maximum, at points where x is $\lambda/2$, λ, $3\lambda/2$ etc. The following formula obtains the resulting stationary wave:

$$p(x,t) = 2p_{max} \cos\left(\frac{2\pi}{\lambda}x\right) \cos(\omega t) \tag{1.1}$$

Stationary waves occur in closed tubes when the tube length d is a full multiple of $\lambda/2$. They are formed at frequencies called *resonance frequencies* or *proper modes*. Assuming there is a maximum sound pressure at the ends of the tube of length L, we have that:

$$\left|\cos\left(\frac{2\pi}{\lambda}x\right)\right| = 1 \text{ in } x = 0 \text{ and } x = L \tag{1.2}$$

For x=L the result is:

$$\left|\cos(\frac{2\pi}{\lambda}L)\right| = 1 \Rightarrow \frac{2\pi}{\lambda} = n\pi \Rightarrow \lambda_n = \frac{2L}{n} \Rightarrow f_n = \frac{cn}{2L} \; n\text{=1,2,3,..} \tag{1.3}$$

Modal frequencies f_n are called fundamental frequency harmonics f_1.
The situation changes if we refer to a three-dimensional environment, such as a parallelepiped. In this case, in the presence of perfectly reflective surfaces, the resulting sound pressure is given by:

$$p(x,y,z,t) = A\cos\left(\frac{2\pi f_x}{c}x\right)\cos\left(\frac{2\pi f_y}{c}y\right)\cos\left(\frac{2\pi f_z}{c}z\right)\cos(2\pi f t) \tag{1.4}$$

where $f = \sqrt{f_x^2 + f_y^2 + f_z^2}$, A is the amplitude and $\lambda = c/f$.
In the hypothesis of an omnidirectional sound source at an angle, the total frequency resulting from the three components x,y,z is obtained and is called the natural resonance frequency f_n.

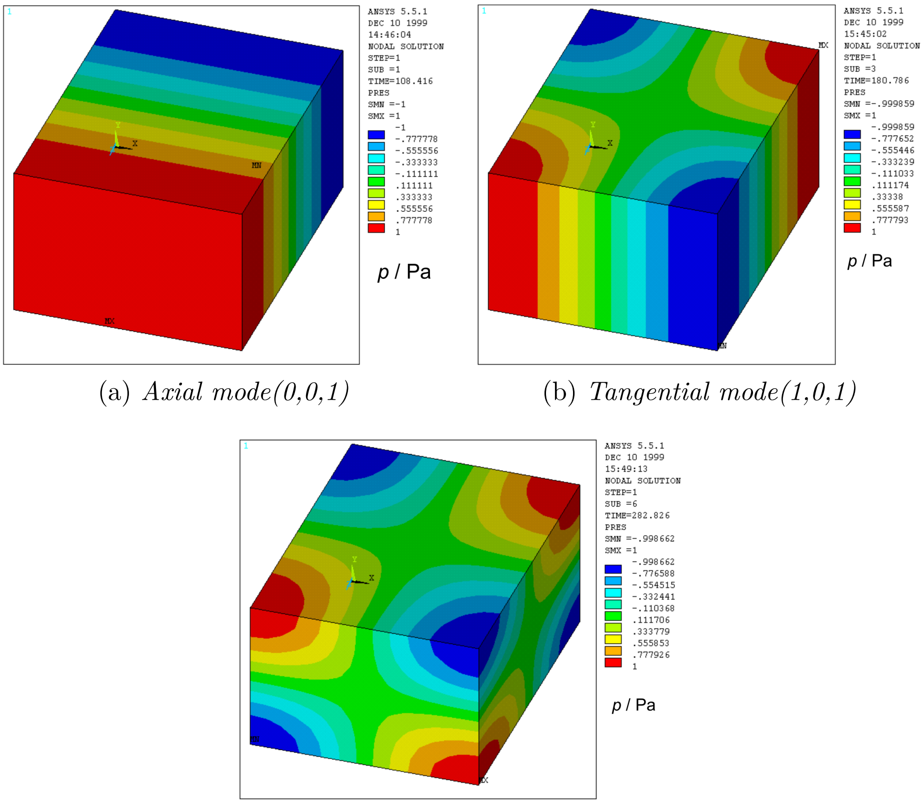
$$f_n = \sqrt{f_x^2 + f_y^2 + f_z^2} = \frac{c}{2}\sqrt{\frac{{n_x}^2}{L_x} + \frac{{n_y}^2}{L_y} + \frac{{n_z}^2}{L_z}} \; n_x, n_y, n_z\text{=1,2,3,..} \tag{1.5}$$

Depending on the three modal indices' values n_x, n_y, n_z three types of modes, or resonance frequencies, are defined.

- **Axial modes:** they depend on a single coordinate and are one-dimensional modes. The waves produced by the omnidirectional source travel parallel to one of the three axes and interact with the pair of surfaces opposite and orthogonal to the axis itself. Moving along a coordinate, the sound pressure changes.

- **Tangential modes:** they depend on two coordinates and are two-dimensional modes. The waves produced by the omnidirectional source travel parallel to the planes related to the two coordinates, interact with the opposite surfaces, and orthogonal to the plane under consideration. Along with two coordinates, the sound pressure changes.

- **Oblique modes:** they depend on three coordinates and are three-dimensional modes. The waves produced by the omnidirectional source propagate in an oblique direction in relation to the three axes interacting with all the pairs of opposing surfaces of the parallelepiped. [7]



(a) *Axial mode(0,0,1)* (b) *Tangential mode(1,0,1)*

(c) *Oblique mode(1,1,1)*

Figure 1.9: Resonance frequencies (modes).

Axial modes have average energy twice as high as tangential modes. Oblique modes are those that have lower average energy; 6 dB less than axial modes.
Therefore, modal resonances create a non-optimal listening experience, creating areas where there are peaks in frequency response and areas where the response is zero. The aim is to make the modal acoustic field as uniform as possible in space and frequency. It is essential to reduce the parallelism of the walls and avoid regular plants, to obtain fewer axial modes in favor of less regular and less energetic modes. When this does not happen, it is necessary to act by acoustically treating the room with materials that do not allow the development of these stationary waves. This work has also been done for the Audio Space Lab.
In the next section, the current features will be reported. For more information on modal response calculations and the reason for the treatments carried out, consult the previous thesis work [4].

1.3.3 Characteristics of the Audio Space Lab

To make the sound field more uniform, the Audio Space Lab has been treated acoustically. As explained later, it is still strong the presence of nodes and antinodes. However, it is a predictable situation because it is an environment similar to a parallelepiped, with equal height and width. The different areas have the

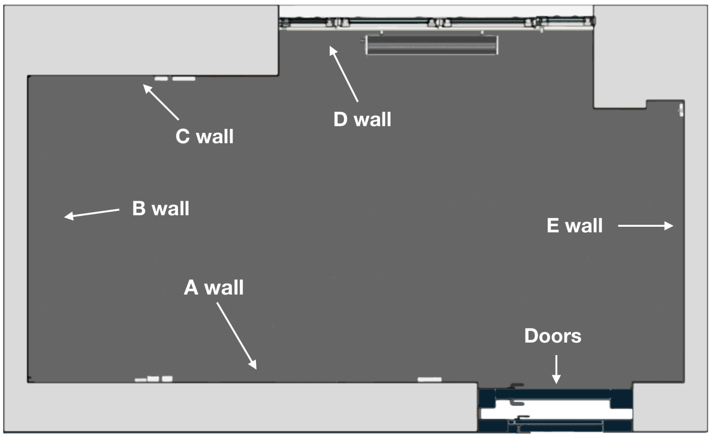


Figure 1.10: Top view of ASL areas..

following characteristics:

- Wall E: counter wall made of plasterboard (thickness 1:25 cm, density 750 kg=m), with an interspace of 6 cm to eliminate the resonance frequency of 76 Hz.
- Walls A, B, C: partially covered with soundproofing panels type Ecophon Akustico^{TM} One SQ2 made up of high density glass wool, a dimension of 600x600 mm, thickness of 40 mm (total thickness, including the anchoring system, is 48 mm) and are positioned in contact with the wall by the Connect One Trim system.
- Wall D: there is a glass window of 2.66x1:77 m.
- Ceiling: covered with sound-absorbing panels of the Ecophon Master^{TM} B Type1 made of an high density glass wool. They have a dimension of 600x600 mm, a thickness of 40 mm (total thickness, including the anchoring system, is 43 mm) and are positioned in contact with the ceiling by glue or dowels.
- Floor: entirely covered with carpet with an average absorption coefficient between 0:25 kHz and 2 kHz equal to d = 0:25.
- External door: soundproofing door with a soundproofing value of R_w=55 dB.

- Internal door: soundproofing door with a soundproofing value of R_w=47 dB.

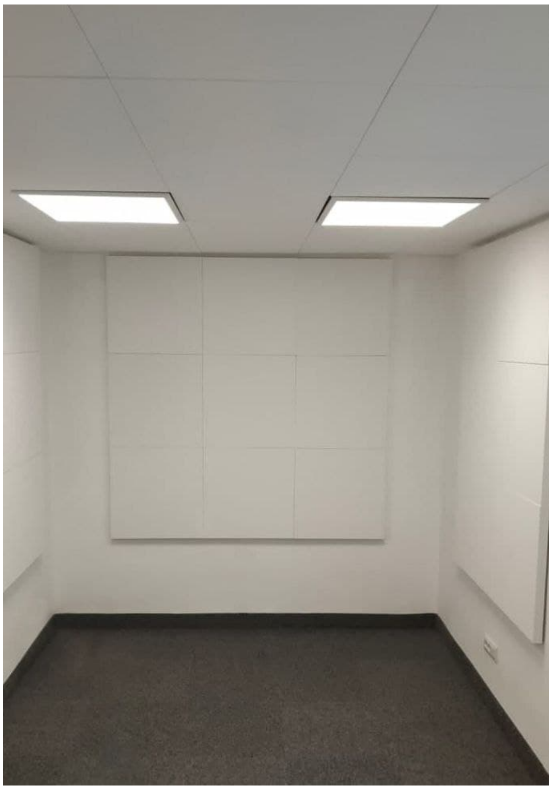


Figure 1.11: Shot from inside the Audio Space Lab.

Figure 1.12 shows the size of the room, which has a volume of 33.3 m^3. The TV (circled in red) is located close to the B wall. The green point indicates the listening position in which the user is seated. He is about two meters from the front wall and at the height of 1.20 m from the floor. Measurements made in the previous study [4] showed different modal frequencies, up to the Schroeder frequency[2] at approximately 134 Hz, and it was verified that the ITU-R BS.1116-3[2] standard criteria were met. In order to have a uniform distribution of acoustic resonances, the following relationship must be satisfied:

$$1.1 \cdot \frac{w}{h} \leq \frac{l}{h} \leq 4.5 \cdot \frac{w}{h} - 4 \tag{1.6}$$

[2] The Schroeder Frequency f_s represents the lower limit of the frequency region in which the sound field assumes statistical connotations (diffuse field); this is the case of large environments. For frequencies smaller than f_s there is a modal behavior.

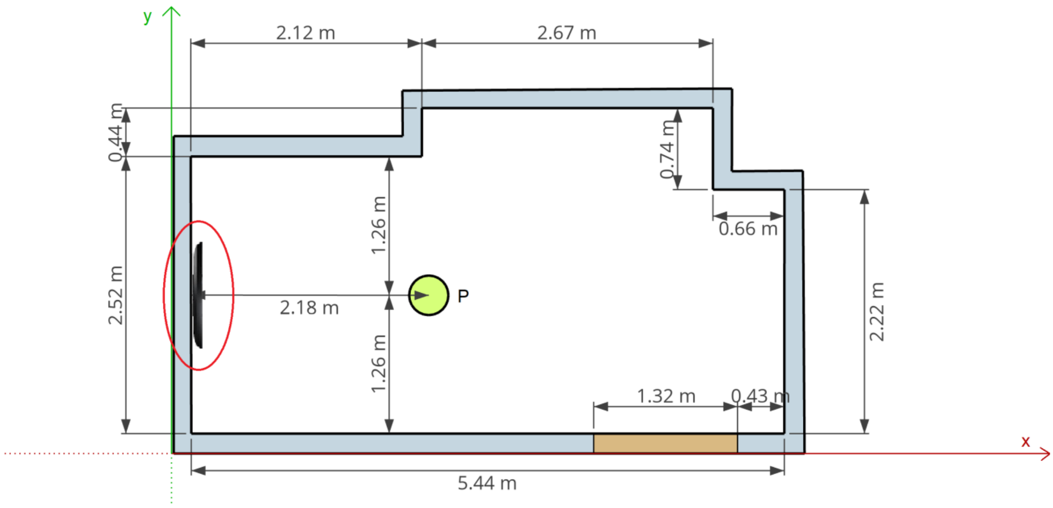


Figure 1.12: Dimensions of the Audio Space Lab.

where l is the length, w the width and h the height, all expressed in meters [m]. Substituting the values, equation 1.6 becomes:

$$1.1 \cdot \frac{2.5}{2.4} \leq \frac{5.4}{2.4} \leq 4.5 \cdot \frac{2.5}{2.4} - 4 \tag{1.7}$$

$$1.14 \leq 2.25 \leq 0.68 \tag{1.8}$$

The relation 1.8 is not met, but the standard specifies that in this case, other conditions have to be verified regarding the sound field. Background noise, reverberation time and listening position requirements have been fulfilled, given the specificity of the test setup, in which the loudspeakers are embedded in the TV-set and the TV-set location has been chosen close to a wall, as in a living room. The TV emits sounds in a frequency range up from 95 Hz (it will be exposed in the following sections), so half of the modal frequencies measured and exposed in Tab. 1.1 do not affect the listening point (Figure 1.13). The most worrying and most energetic is the 127.8 Hz axial mode along the x-axis, so the listener has been slightly moved so that it is not at a node or anti-node.

The ITU-R BS.1116-3 standard defines that the mean reverberation time T_m measured in the range of 200 Hz to 4000 Hz should be:

$$T_m = 0.25 \left(\frac{V}{V_0} \right)^{1/3} \tag{1.9}$$

where V is the volume of the room in m^3 and V_0 is a reference volume established at 100 m^3. The tolerance to be applied to the range 250 Hz to 4 kHz is $\pm$ 0.05 s, and as shown in the Figure 1.14 meets the requirement.

Mode	(nx; ny; nz)	Average in listening position
Axial	(1,0,0)	32.3 Hz
Axial	(2,0,0)	56.5 Hz
Axial	(0,1,0)	64.6 Hz
Axial	(0,0,1)	72.6 Hz
Tangential	(1,1,0)	–
Tangential	(1,0,1)	–
Tangential	(2,1,0)	91.5 Hz
Axial	(3,0,0)	–
Tangential	(2,0,1)	94.8 Hz
Tangential	(0,1,1)	–
Oblique	(1,1,1)	101.6 Hz
Tangential	(3,1,0)	116.4 Hz
Oblique	(2,1,1)	–
Tangential	(3,0,1)	123.8 Hz
Axial	(4,0,0)	127.8 Hz
Axial	(0,2,0)	–
Oblique	(3,1,1)	–
Tangential	(1,2,0)	138.6 Hz

Table 1.1: Normal modes of the Audio Space Lab, average in the listening position.

Inside the room, there is a background noise produced by the air conditioning system, various equipment, and other external sources. The standard states that it should fall between curve NR[3] 10 and NR 15. Figure 1.15 shows the measurement taken at a height of 1.20 m at the listening point; guidelines are respected, except in the octave bands from 2 kHz to 8 kHz.

[3] *Noise Rating* curves have been developed to determine whether the indoor environment is acceptable for hearing conservation, speech communication and annoyance. There are different acceptable sound pressure levels that vary depending on the type of environment and its use, so there are many different curves, from NR 0 to NR 130.

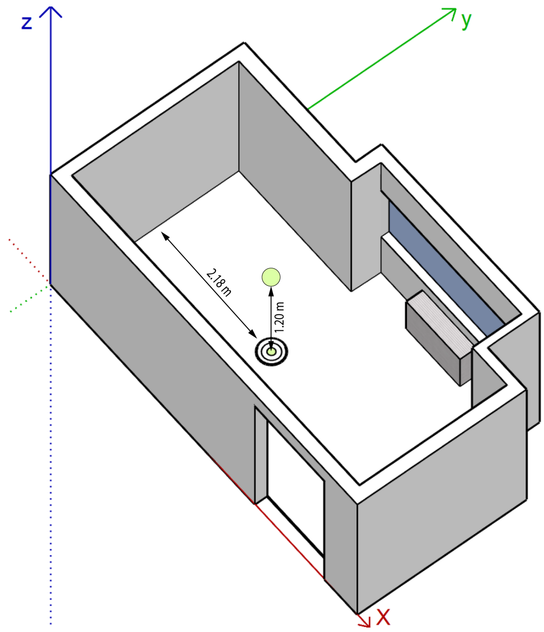


Figure 1.13: Position of the listening point P.

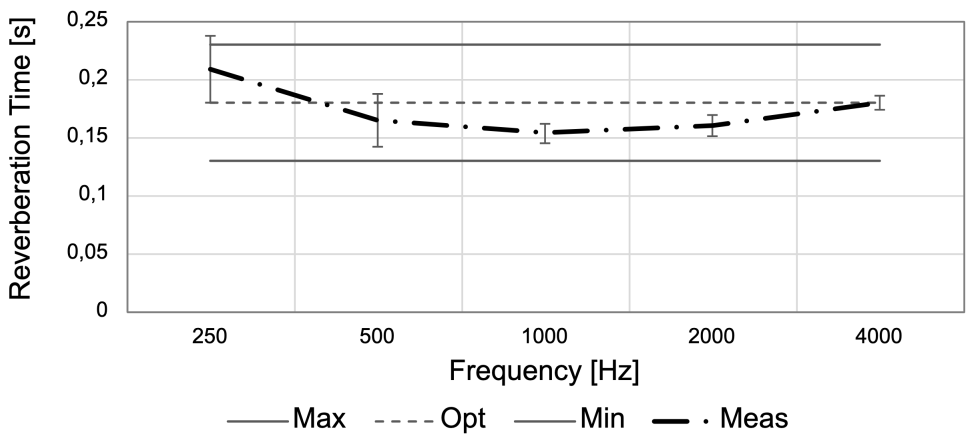


Figure 1.14: Measured reverberation time of the ASL and tolerances according to the standard.

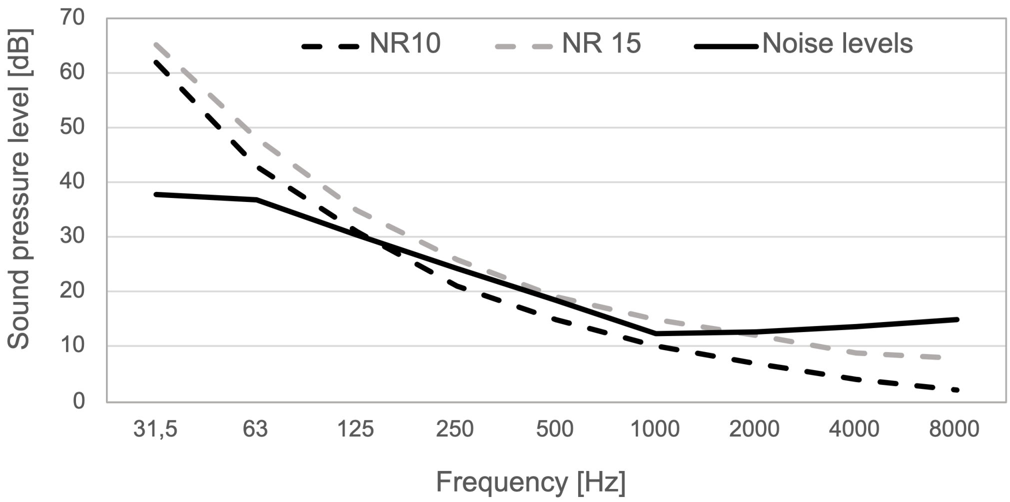


Figure 1.15: Measured background Noise Level.

Chapter 2

Test listening arrangement

This section explains the TV's characteristics by analyzing its impulse response conducted in the previous study[4], to observe any criticality in the spectrum that does not allow the reproduction of specific frequencies. To this end, two measurements were made: the first in an anechoic room, the second in a non-anechoic environment, namely the Audio Space Lab. The characteristics of the material used for the listening test are also displayed. These are short-length audio/video tracks provided by RAI and taken from programs of their daily schedule. ITU-R BS.1116-3 has been used as a guideline to achieve the best possible working conditionsm even for subsequent subjective tests.

2.1 TV and user positioning

Figure 2.1 shows the configuration recommended by the ITU R BS.1116-3 standard. The setup shown is that of a stereophonic configuration in which The listening position is placed at the center of the circle of diameter D, respectively 30° from the left(L) and right(R) speakers. Unfortunately, it is not possible to comply with these conditions, although the TV has a stereo output, as the two speakers are fixed without the possibility of movement. The standard indicates that distance B can take a value of 2-3 m and D a value (1.7-2)B. Based on this, the listening position was positioned 2 meters from the TV (Figure 2.2) and at a height from the ground of about 1.2 m; This last one corresponds to the average height of a seated listener's head.

2.2 Impulse response measurements

Two measurements were carried out. The first one at the anechoic room to analyze the TV response without any reflection, and the second one at the Audio

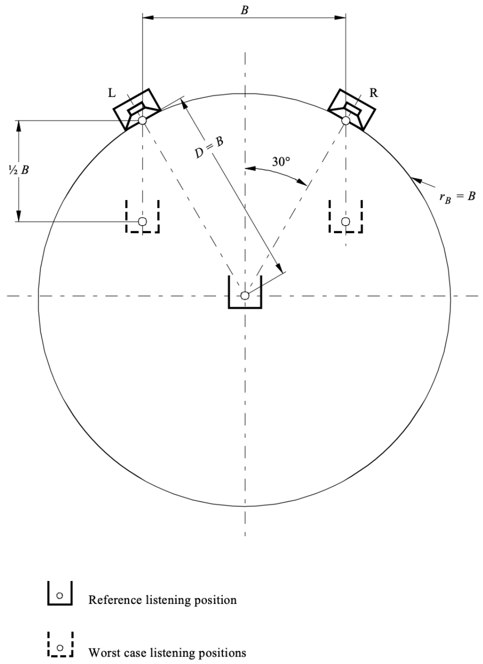


Figure 2.1: Setup reported in the IT R BS.1116-1 standard.

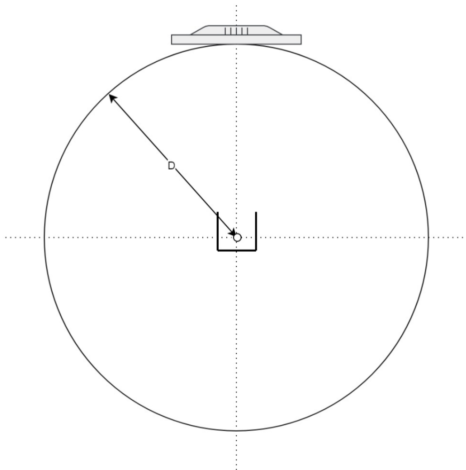


Figure 2.2: Test listening arrangement in this project.

Space Lab to evaluate the room's response in which the user performs the test. Measurements were made using the Sine Sweep method. This method consists of sending a sweep to the TV and convolving the recorded signal with the inverse of the sweep. A sweep consists of a sine wave with a constant amplitude, the frequency of which varies exponentially with time. The need to reverse the sweep is due to the fact that any signal mixed with its inverse time produces a pulse. The convolution between the recorded signal and the inverse of the original allows the calculation of the room's impulse response. In the datasheet, the supplier declares that the frequencies emitted from the television go from 100 Hz to 8000 Hz. This information was used to decide the starting and endpoint of the sweep to be generated.

2.2.1 TV frequency response in anechoic chamber

The measurements shown in Figure 2.3 allows to visualize the spectrum coming directly from the TV without other influences. The measurements shown in Figure 2.3 allow visualizing the spectrum coming directly from the TV without other influences. The spectrum is more or less flat from 100 Hz to about 1400 Hz, with a mean value of (-50±10) dB. It is possible to notice a hole about 80 Hz wide in the zone of 1550 Hz. With increasing frequency, there is a decay of sound pressure of -60 dB/dec.

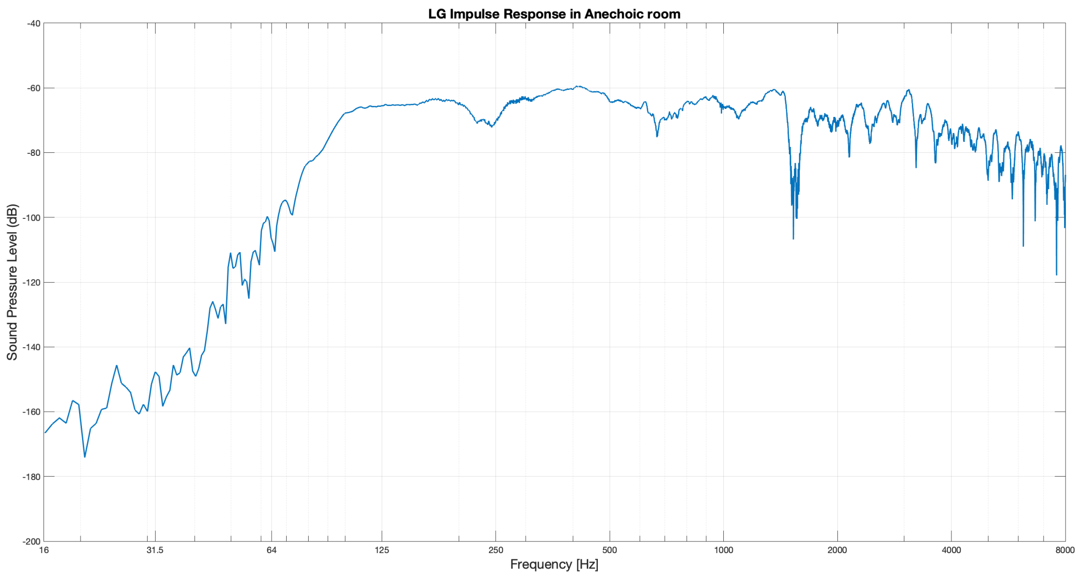


Figure 2.3: TV impulse response in Anechoic room.

2.2.2 TV frequency response in Audio Space Lab

Figure 2.4 displays the spectrum of measured frequency response at the Audio Space Lab. Between 100 Hz and 1000 Hz the sound pressure level is on average (-70±15) dB with an area that drops to -90 dB in the frequencies of 200 Hz.

Starting at 1000 Hz, the spectrum is more jagged with holes up to about 40 dB. There is a general decay of -7 dB/oct.

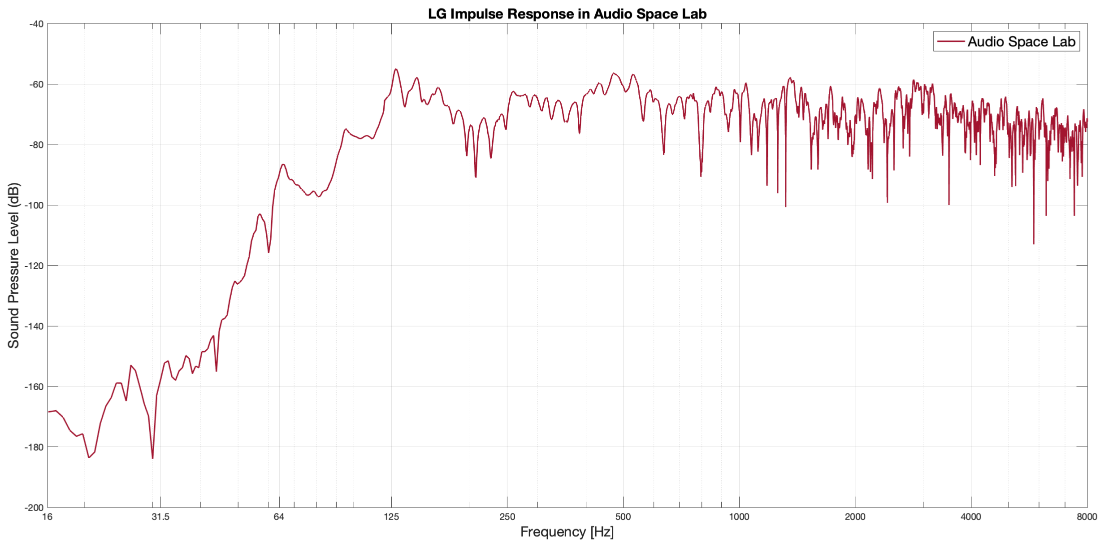


Figure 2.4: TV impulse response in Audio Space Lab.

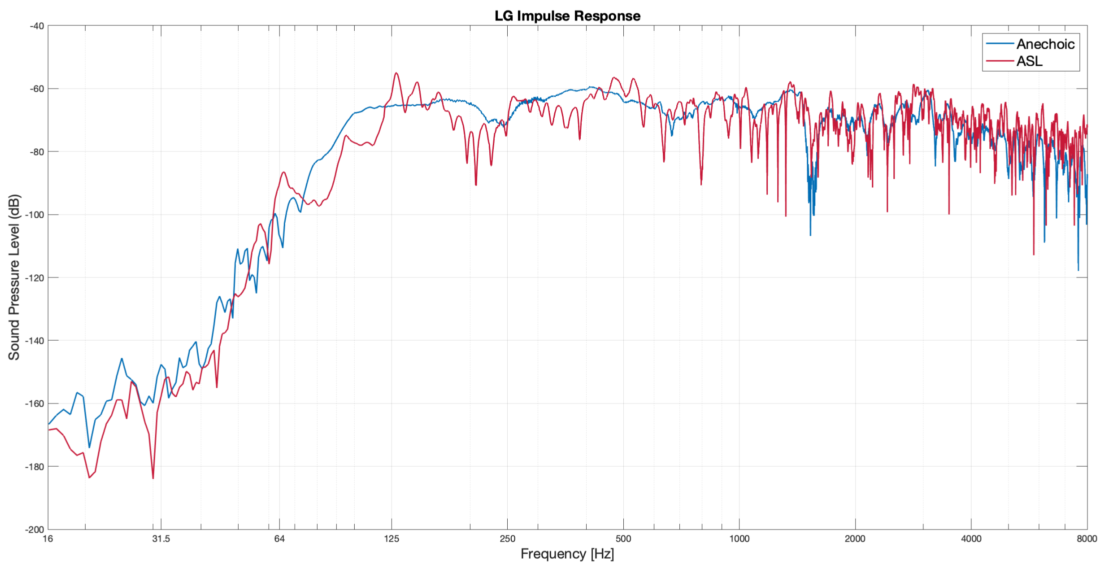


Figure 2.5: Impulse response in Anechoic room and Audio Space Lab.

2.3 Selection of video extracts

The material employed during this research activity is the same as the previous study since the two projects share the same goal: improving speech intelligibility. RAI provided a series of video fragments extracted from the schedule's programs and aired between 2017 and 2019. The extracts have been divided into three categories to identify the characteristics that unite the different genres and then analyze the spectrum for the choice of an optimal transfer function. The three different genres are:

- **Sport**: it includes news reports of sporting events, such as football, basketball and volleyball. In these excerpts, the voice of the commentator is influenced by the background noise of the stadium or the external environment in which it is located.
- **Speech**: it includes films, news, television fiction, and documentaries. In this kind of excerpt, the dialogues are well distinct from the context, and no background noise or music influences them.
- **Music**: it covers music programs such as talent shows or singing shows and the like. These extracts are contradistinguished by sung dialogues and a musical context that interacts with the frequencies of speech.

The ITU-R BS.1116-3 [2] standard stipulates that for this kind of test, there must be at least five extracts per genus of duration between 10 s and 25 s. 30 extracts were chosen about 10 s long each, distributed in this way:

- 18 tracks for *Speech* genre
- 5 tracks for the *Music* genre
- 7 tracks for the *Sport* genre

Frequency analysis of genres

Through the software Audition[8] was analyzed the frequency domain of each track to create a spectrum that enclosed the characteristics of each of the three genres chosen. The results were analyzed to assess areas with common characteristics and improve the spectrum that applied to all three genera. Figure 2.7 shows a similar behavior (especially for Speech and Sport) in the frequency bands higher than about 600 Hz, with a slope of about -20 dB/oct, so the changes that will be made to improve intelligibility will affect all tracks.

(a) *Example of a "Sport" program*

(b) *Example of a "Music" program.*

(c) *Example of a "Speech" program.*

Figure 2.6: Examples of RAI programmes.

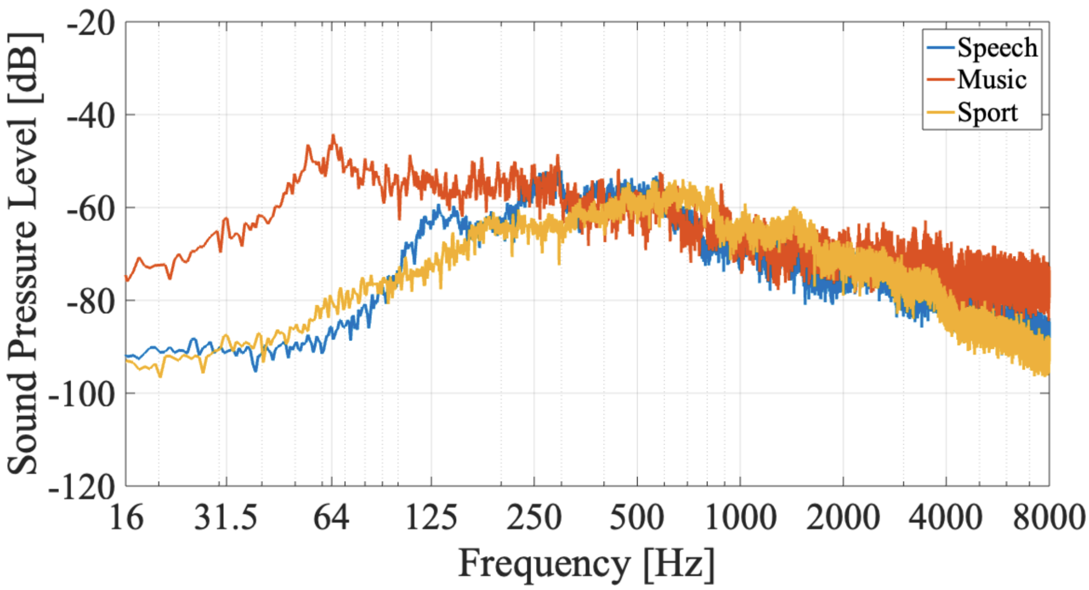


Figure 2.7: Average spectra for each genre.

Chapter 3

DSP Setup and Transfer functions used

This chapter deals with speech intelligibility by considering the frequency bands that characterize it. Furthermore, the two transfer functions (t.f.) employed in this thesis are described. The first one is the "Heavy" t.f., which was used in the previous research project and the second is the "Moderate" t.f. developed working with otolaryngologists of the University of Turin.

3.1 Speech intelligibility

Intelligibility is the percentage of words or understood sentences on the totality of those pronounced during verbal communication. Intelligibility shall be assessed based on:

- *signal level* which depends on the human voice's acoustic characteristics (power, directivity, frequency composition) and the distance between speaker and listener.
- *level of noise* determined by presence or absence of background noise and the distance between listening and a particular source of the noise.

3.1.1 Speech Transmission Index (STI)

There are several indices used for intelligibility analysis. One of these is the Speech Transmission Index (STI) which is based on the concept of carrier modulation, as it is assumed that human speech can be represented in this way. In the classical STI method, the carrier is represented by a stationary Gaussian signal divided into seven octave bands of amplitude equal to 1/2 octave band in the 125 Hz - 8000 Hz

range. Moreover, each band is modulated in amplitude with a frequency variable from 0.63 at 12.5 Hz, with a 1/3 octave pitch (total 14 modulation frequencies). So it is a complex signal consisting of a set of 98 combinations $\boldsymbol{m}$ (14 modulation frequencies x 7 frequency bands), which must be reproduced by a driver unit with human mouth dimensions. The values $\boldsymbol{m}$ are corrected to consider the effects of auditory masking and the absolute threshold of reception. These values are converted into apparent SNR($L_s - L_n$) signal-noise ratios, limited to the range of ± 12 dB:

$$SNR_{f,F} = 10log\frac{m_{f,F}}{1 - m_{f,F}}\ [dB] \tag{3.1}$$

Each signal-noise apparent ratio is converted to TI$_{f,F}$ transmission index, in the range of 0 to 1:

$$TI_{f,F} = \frac{SNR_{f,F} + 15}{30} \tag{3.2}$$

The MTI$_f$ modulation transfer index is calculated for each octave band:

$$MTI_f = \frac{1}{14}\sum_{F=1}^{1} 4TI_{f,F} \tag{3.3}$$

The STI index is obtained as the weighted sum of modulation transfer indices for all 7 octave bands:

$$STI = \sum_{f=1}^{7} \alpha_f MTI_f - \sum_{f=1}^{6} \beta_f \sqrt{MTI_f \cdot MTI_{f+1}} \tag{3.4}$$

α_f and β_f are the correction coefficients for the different octave bands, and they are different in relation to the speaker gender. The first is a weight factor of the octave band, that expresses the relative contribution of each octave band to the STI. The second is a redundancy factor of the octave band (Table 3.1).

Gender of the talker		**Central frequencies of the octave bands (Hz)**						
		125	**250**	**500**	**1000**	**2000**	**4000**	**8000**
Male	α	0.085	0.127	0.230	0.233	0.309	0.244	0.173
	β	0.085	0.078	0.065	0.011	0.047	0.095	-
Female	α	-	0.117	0.223	0.216	0.328	0.250	0.194
	β	-	0.099	0.066	0.062	0.025	0.076	-

Table 3.1: Factors representing the relative contribution of each octave band to the TSI index in accordance with IEC 60268-11.

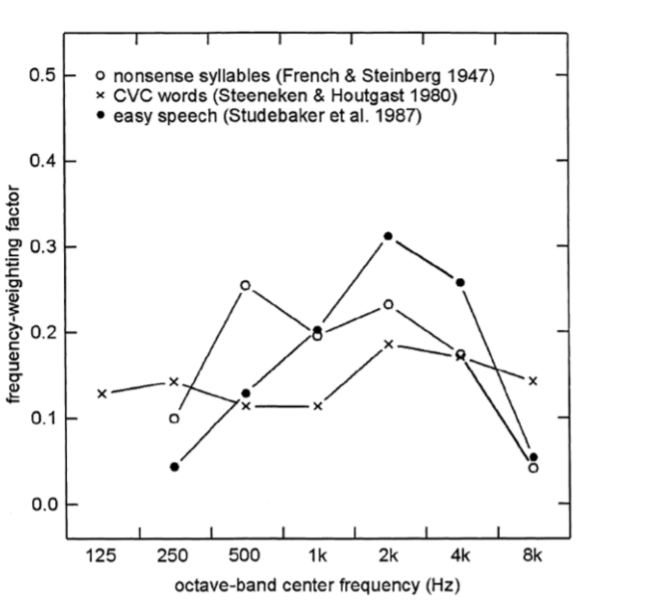


Figure 3.1: Bandwidth weighting factors for intelligibility contribution, in octave bands.

Looking at Figure 3.1, the most interesting frequencies for good intelligibility are those between 500 Hz and 4000 Hz. Tables 3.2 and 3.3 show the band weighting factors for the contribution of intelligibility, relative to different evaluation tests. The first table displays contributions in octave bands, the second in 1/3 octave bands [9].

Band No.	Midband Freq, Hz	NNS[a]	CID-22[b]	NU6[c]	DRT[d]	Short Passages[e]	SPIN[f]
1	250	0.0437	0.1549	0.0853	0.0960	0.1004	0.0871
2	500	0.1294	0.1562	0.1912	0.2043	0.2551	0.1493
3	1000	0.2025	0.2165	0.2110	0.2343	0.1960	0.2206
4	2000	0.3117	0.2768	0.3090	0.2643	0.2322	0.3022
5	4000	0.2576	0.1488	0.1682	0.1501	0.1744	0.2102
6	8000	0.0551	0.0468	0.0353	0.0510	0.0419	0.0306

[a]NNS (various nonsense syllable tests where most of the English phonemes occur equally often), [b]CID-W22 (PB-words), [c]NU6 monosyllables, [d]DRT (Diagnostic Rhyme Test), [e]short passages of easy reading material, [f]SPIN monosyllables

Table 3.2: Octave band importance functions for various speech tests [9]

The speech undergoes continuous changes in the sound level related to the language. In normal voice conversation and excluding silence, the voice dynamic is

Band No.	Midband Freq, Hz	NNS[a]	CID-22[b]	NU6[c]	DRT[d]	Short Passages[e]	SPIN[f]
1	160	0.0000	0.0365	0.0168	0.0000	0.0114	0.0000
2	200	0.0000	0.0279	0.0130	0.0240	0.0153	0.0255
3	250	0.0153	0.0405	0.0211	0.0330	0.0179	0.0256
4	315	0.0284	0.0500	0.0344	0.0390	0.0558	0.0360
5	400	0.0363	0.0530	0.0517	0.0571	0.0898	0.0362
6	500	0.0422	0.0518	0.0737	0.0691	0.0944	0.0514
7	630	0.0509	0.0514	0.0658	0.0781	0.0709	0.0616
8	800	0.0584	0.0575	0.0644	0.0751	0.0660	0.0770
9	1000	0.0667	0.0717	0.0664	0.0781	0.0628	0.0718
10	1250	0.0774	0.0873	0.0802	0.0811	0.0672	0.0718
11	1600	0.0893	0.0902	0.0987	0.0961	0.0747	0.1075
12	2000	0.1104	0.0938	0.1171	0.0901	0.0755	0.0921
13	2500	0.1120	0.0928	0.0932	0.0781	0.0820	0.1026
14	3150	0.0981	0.0678	0.0783	0.0691	0.0808	0.0922
15	4000	0.0867	0.0498	0.0562	0.0480	0.0483	0.0719
16	5000	0.0728	0.0312	0.0337	0.0330	0.0453	0.0461
17	6300	0.0551	0.0215	0.0177	0.0270	0.0274	0.0306
18	8000	0.0000	0.0253	0.0176	0.0240	0.0145	0.0000

[a]NNS (various nonsense syllable tests where most of the English phonemes occur equally often), [b]CID-W22 (PB-words), [c]NU6 monosyllables, [d]DRT (Diagnostic Rhyme Test), [e]short passages of easy reading material, [f]SPIN monosyllables

Table 3.3: One-third octave band importance functions for various speech tests [9]

about 30 dB, expressed as the difference between the maximum value and the minimum value of the emitted level. Vowel sounds have a louder and lower frequency level and last longer than consonant sounds, which can also be mute or impulsive.

3.2 Presbycusis

Presbycusis reduces the auditory capacity with age for physiological aging phenomena with gradual onset and slow progression. The presbycusis is characterized by a bilateral neurosensory hypoacusia, symmetric that begins mainly affecting the acute frequencies (over 4000 Hz). Only after the age of 65, the hearing loss begins to interest the language's typical frequencies. One of the characteristics of presbycusis is the decrease in the ability to recognize the word. Presbycusis is a phenomenon linked exclusively to aging. However, it must also be remembered that, even with the advancing age, there may be several factors of alteration of the auditory system. The main ones are:

- The thickening of the tympanic membrane;
- The cellular degeneration of the organ of the Corti, an organ that transmits sound impulses at the central level;

- A reduced elasticity of the cochlea's basic membrane;
- A smaller number of auditory eyelashes;
- Compression of nerve fibers due to hyperostosis phenomena.

Presbycusis can be diagnosed clinically through audiometric examination, although the symptoms are quite obvious. This test reveals that an affected subject has a threshold without discrepancies in lower frequencies, and instead an elevation in the area of high frequencies.
There is no medical or surgical therapy that can restore the lost auditory capacity. It may be useful to implement good prevention by avoiding exposure to factors that may damage the auditory system, in particular prolonged exposure to noise and ototoxic substances. The application of binaural prostheses, which amplify above all the medium-acute tones, without excessively closing the external auditory conduit, is able to return a part of the lost frequency spectrum.

3.3 "Heavy" Transfer Function

The "Heavy" transfer function was developed during the previous study[4] and for its implementation was considered that:

1. The most important frequencies for speech intelligibility are those between 500 Hz and 4000 Hz [10][11][12];
2. The human ear is most sensitive in the range from 3 kHz to 4 kHz;
3. The TV frequency spectrum has a decay from about 2 kHz, as Figure 2.5 shows;
4. The tracks used (the same as in this study) have a similar behaviour in the range from about 600 Hz and a decay of -20 dB/oct.

Figure 3.2 shows the "Heavy" transfer function. The white line represents the threshold, created by researchers and set on DSP, used to compare the input signal and bring it to the levels indicated by the threshold itself. The circles correspond to the maximum amplification or attenuation that can be applied by the software to each given band. As mentioned, this equalization is dynamic and changes depending on the track's frequency content, so as to keep the signal energy constant. The values for each circle are shown in Tab. 3.4.

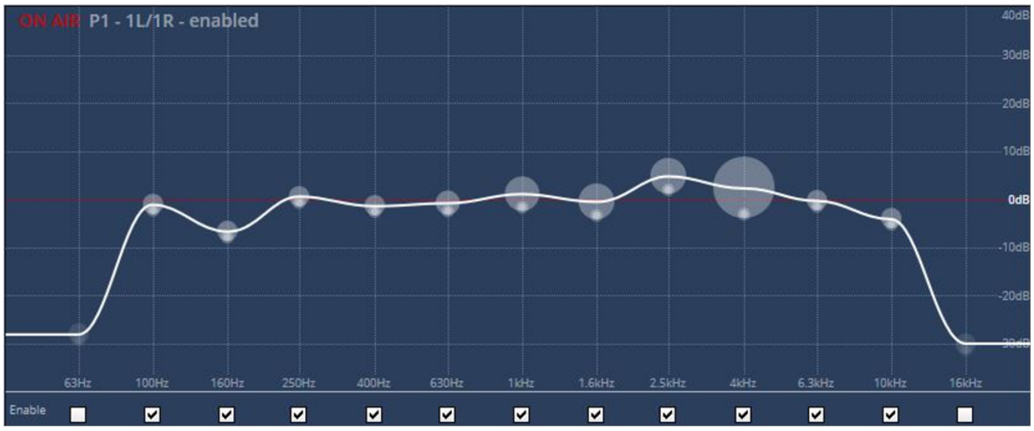


Figure 3.2: "Heavy" Transfer Function.

Frequency $[Hz]$	**Max Gain** $[dB]$	**Threshold level** $[dB]$
100	2	-1.1
160	1.6	-6.7
250	1.5	0.6
400	2.2	-1.4
630	3.7	-0.8
1000	5.8	1.1
1600	6.1	-0.5
2500	6.1	4.8
4000	12.0	2.3
6300	1.1	-0.3
10000	0.7	-4.1

Table 3.4: Band settings for "Heavy" transfer function.

3.4 "Moderate" Transfer Function

The transfer function "Moderate" was implemented with the help of otolaryngologists of the University of Turin on the basis of the following aspects:

1. Human physiology studies report a maximum resonance of the external auditory canal around 2700 Hz and of the middle ear between 800 and 1500 Hz[13];

2. The frequency range around 2 kHz is the most important one for intelligibility, intended as the percentage of words correctly understood on the total of the words submitted to the subject[14]. Therefore, it was decided to amplify

more the one-third octave frequency bands with a central frequency of 1.6 kHz and 2 kHz.

3. In the medical-legal field of hearing loss caused by noise, the social value of hearing function was taken as a reference in terms of understanding the voice of conversation. As a result, each frequency was given a different weight, as shown on INAIL Circular Letter No. 22 of 07 July 1994[15]. The percentage weight of frequencies was redistributed on the basis of their utility: 25% for 500 Hz and 1000 Hz, 35% for 2000 Hz, 10% for 3000 Hz and 5% for 4000 Hz.

Figure 3.2 shows the "Heavy" transfer function Figure 3.3.

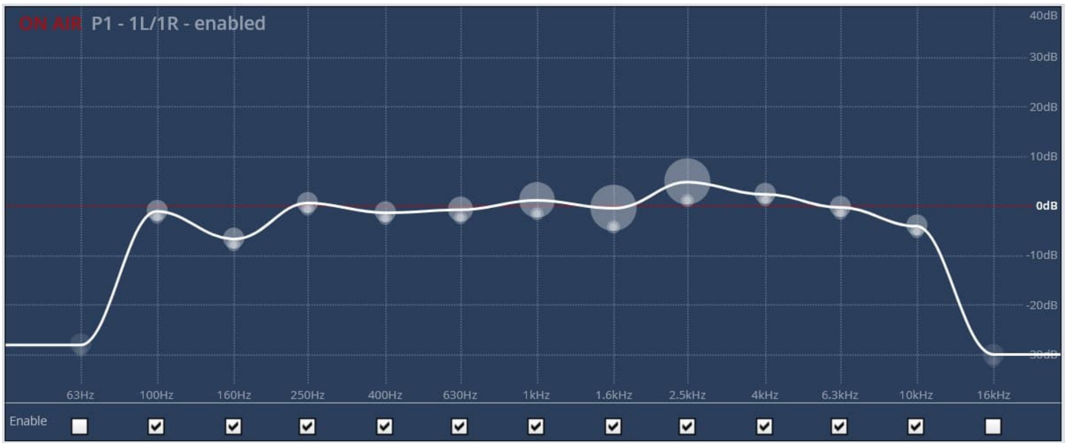


Figure 3.3: "Moderate" Transfer Function.

For maximum gains and the threshold set for the audio processor, the same considerations apply as previously. Tab 3.5 gathers the values.

3.5 Transfer functions during processing

In the following two sections, the two transfer functions are analyzed while working in each of the three chosen genres (Speech, Music, Sport). Note that the "Moderate" filter modifies the signal more narrowly than the "Heavy" filter.
In each image, there are two curves: a *dotted* line that represents the input signal as the DSP receives it and a *continuous* line that represents the output signal modified after processing.

3.5.1 "Heavy" filter processing

The following is an example for each genre chosen:

Frequency $[Hz]$	**Max Gain** $[dB]$	**Threshold level** $[dB]$
100	2	-1.1
160	1.6	-6.7
250	1.5	0.6
400	2.2	-1.4
630	3.7	-0.8
1000	5.8	1.1
1600	8	-0.5
2500	8	4.8
4000	3	2.3
6300	1.1	-0.3
10000	0.7	-4.1

Table 3.5: Band settings for "Heavy" transfer function.

- *Music* track: Figure 3.4 shows the processing carried out on the track of the "Music" genre. It is uniform in the affected frequency bands since audio tracks are very rich in frequency, even in areas not affected by language.

Figure 3.4: "Heavy" filter during music track processing.

- *Speech* track (3.5): It is possible to notice more significant processing of frequencies related to speech. The latter is not affected by other contributions, especially the peak around the 4 kHz characteristic of this transfer function.

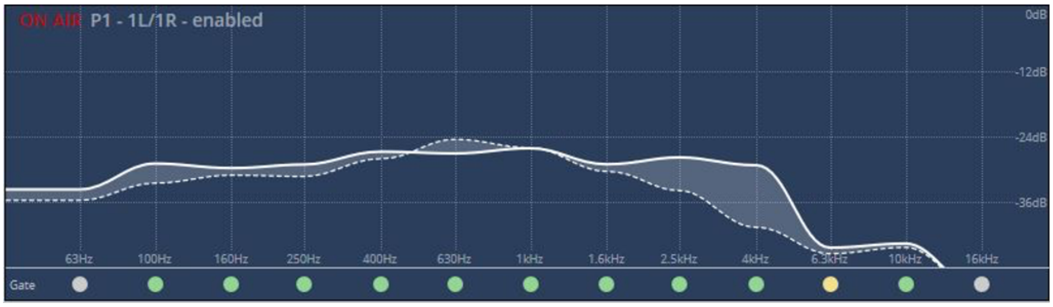


Figure 3.5: "Heavy" filter during speech track processing.

- *Sport* track (Figure 3.6): The tracks related to this category are similar to those in the "Speech" category, so the processing is similar.

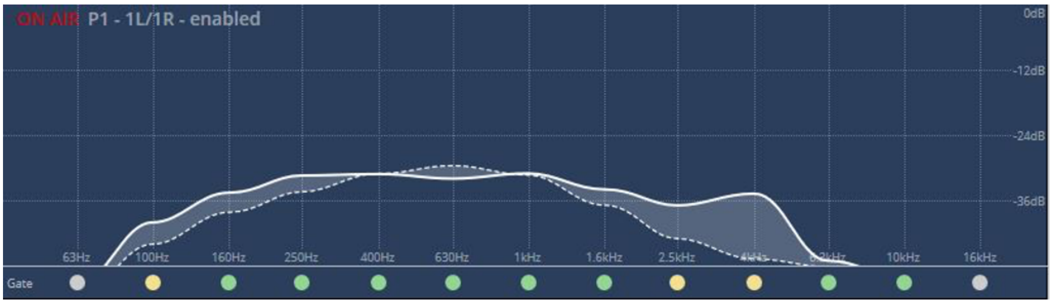


Figure 3.6: "Heavy" filter during sport track processing.

3.5.2 "Moderate" filter processing

Figures 3.7, 3.8, 3.9 show some examples of the "Moderate" transfer function, for each category. This filter performs a lower processing rate, especially around 4 kHz where the difference between the input signal (dotted line) and the output signal (continuous line) is less marked.

3.6 Digital Signal Processor programming

The role of DSP (Junger D*AP4 FLX) is to alter the audio signal coming from RAI programs before being broadcast and reach the user's TV. Audio processing consists of dynamic equalization; i.e., signal processing depends on the frequency content of the input signal and processing adapts in real time to a reference curve (the transfer function, or "Spectral Signature" as it is called on the *Junger* configuration page). Suppose a signal has a sound pressure level in the analyzed band, comparable with the reference curve. In that case, the equalizer does not force

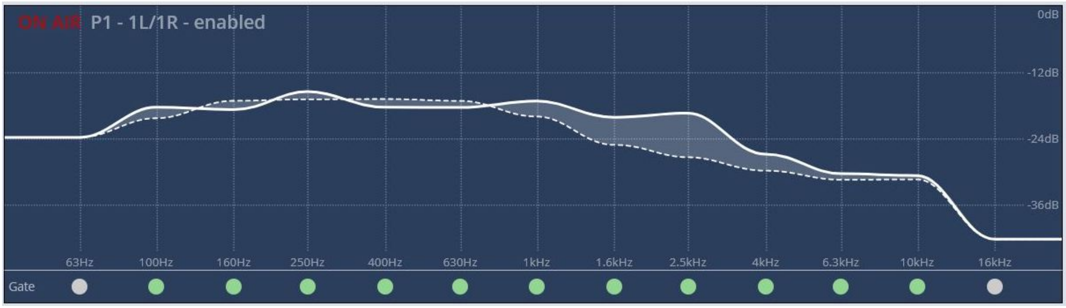


Figure 3.7: "Moderate" filter enabled on a Music track.

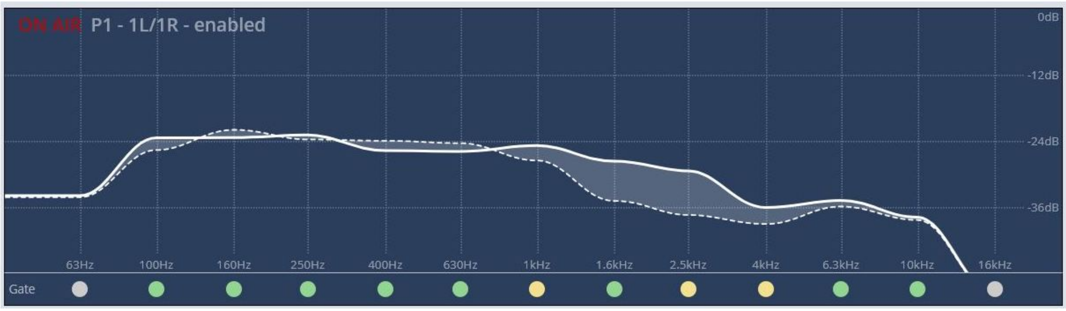


Figure 3.8: "Moderate" filter enabled on a Speech track.

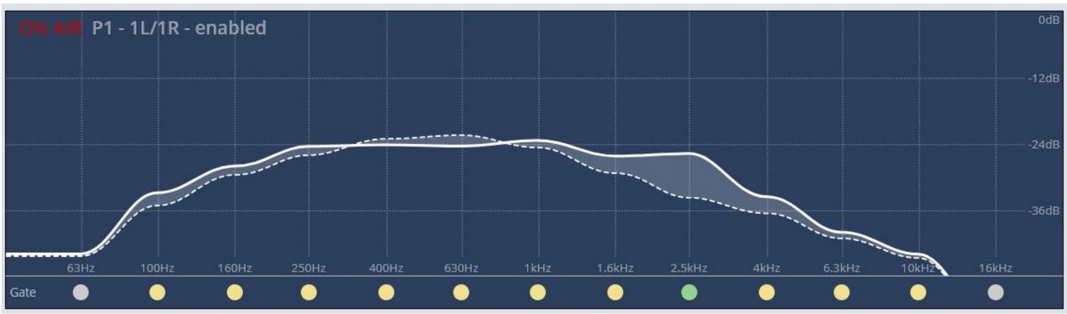


Figure 3.9: "Moderate" filter enabled on a Sport track.

a further increase, which would otherwise cause a not negligible distortion. Conversely, if the sound pressure level in the analyzed band is lower than the reference curve, the DSP will automatically increase it. Thirteen frequency bands were configured, and they correspond to the central frequencies of one-third octave bands, taken one every two. The Junger D*AP4 FLX allows selecting the gain or attenuation that each band can undergo concerning the reference curve in an interval of (0 - 12) dB.

3.6.1 EBU R128 Regulation

The processor dynamically decides how to process the signal based on the information described up to this point. But, above all, respecting the rule regarding the levels of Loudness[1]. In fact, for the programs of the television station, reference is made to the parameter regulated by the EBU R128[3], which says:

> . . . the **Programme Loudness Level** shall be normalised to a Target Level of **-23.0 *LUFS***[2]. Where attaining the Target Level is not achievable practically (for example, live programmes), a tolerance of **±1.0 LU** is permitted. [. . .] and that the **True Peak Level** of a programme shall not exceed **-1 *dB TP*** (*dB True Peak*) during production (linear audio), measured with a meter compliant with ITU-R BS.1770 and EBU Tech 3341.

3.6.2 Settings

The Junger has a static IP address through which it can be programmed from the workstation, via Ethernet. A page opened by the browser allows to set some parameters that will be exposed below.

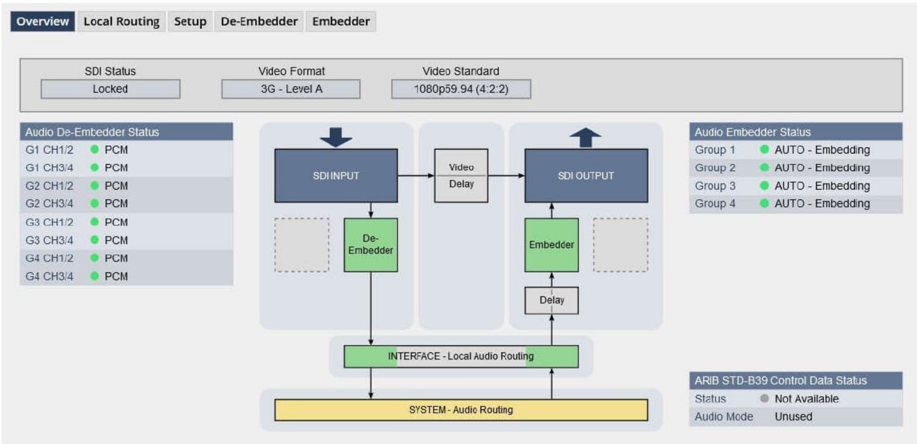


Figure 3.10: Overview of the signal path.

Figure 3.10 shows the path that audio and video signals follow after passing the HDMI-to-SDI converter. The audio signal is processed during various DSP

[1] Loudness is the subjective perception of sound intensity.

[2] *LUFS* stands for Loudness Units relative to Full Scale. It is a standardized measure of the audio volume that calculates together the human perception and the intensity of the electrical signal. The *LUFS* are used to set objectives for the normalization of audio in transmission systems for streaming cinema, TV, radio and music.

stages, while the video one must wait by going through a delay (a fixed time interval of about 6 ms, which takes into account the audio processing time) before being merged again to the audio signal. They then flow through the SDI-to-HDMI converter to be played. Within the audio processor various plugins are available, as shown in Figure 3.11.

In this study, the block related to the Filters was used, where the transfer function

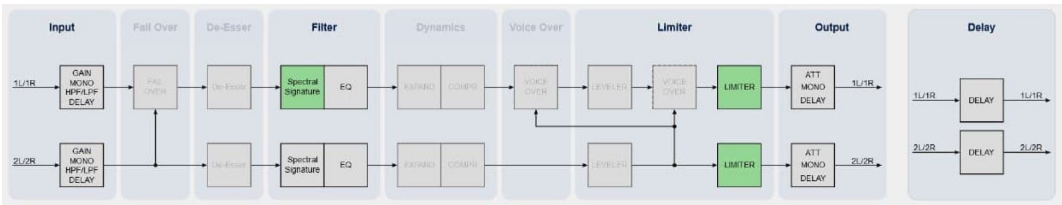


Figure 3.11: Plugins available in the Junger D*AP4 FLX.

was implemented, and that of the Limiter. The latter served to control the peaks that can be generated after the processing of the audio signal. In Figure 3.12 is shown the parameter,"Max True Peak", that regulates this aspect. The last aspect

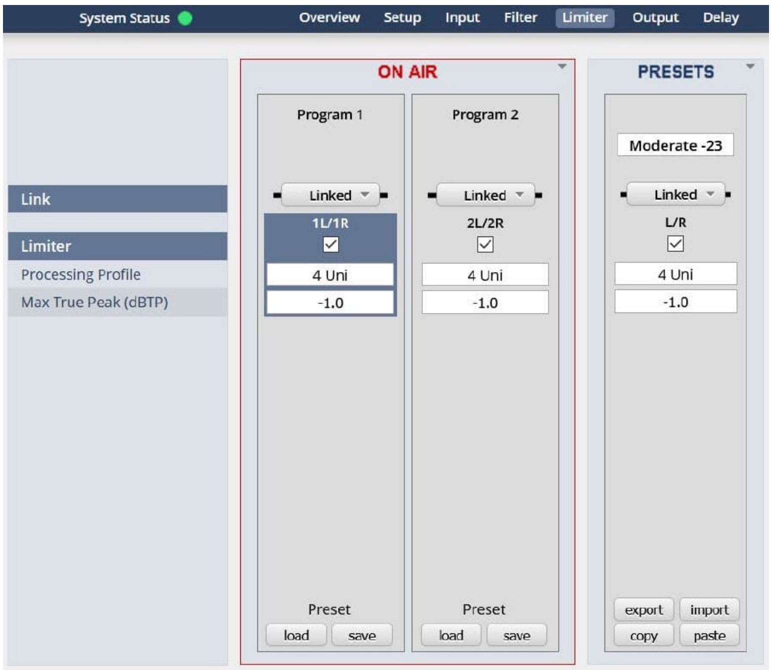


Figure 3.12: Limiter plugin.

is that concerning the EBU R128 regulation, discussed in the previous section. The value of Loudness must not exceed -23 *LUFS* in a certain period of time. It is possible to insert the chosen legislation and monitor the performance. In the frame shown in Figure 3.13, it is possible to observe as an input of -25.6 *LKFS* (Synonym of *LUFS*) corresponds to an output of -23.9 *LKFS*.

Figure 3.13: Loudness measurement.

Part II

Experimental methodology and data analysis

Chapter 4

Sebjective tests

This chapter describes the methodology used to evaluate the transfer functions described above and the subjective tests performed at the Audio Space Lab of the Politecnico di Torino. The selected volunteers were subjected to an audiometric test to verify their auditory threshold and to a listening test identified thanks to the ITU-R BS.1116-3[2] norm. In particular, the method of the "double-blind triple-stimulus with hidden reference" described in the next section was used. The test is divided into three parts:

1. Completion of a written release to be submitted to the test and use the data obtained from the test (5 minutes);
2. Audiometric test (10 minutes);
3. Training test (15 minutes);
4. Main test (30 minutes).

The test participants are divided into two groups: elderly people with presbycusis, without otological pathologies, over 65 years of age, and young non-native speakers. The results obtained for the two groups are presented in the final section.

4.1 Reference recommendation ITU-R BS.1116

The following points are those of the standard used as reference. Also, there are some comments related to this study.

- **Pre-screening of subjects:**

  > "Pre-screening procedures, include methods such as audiometric tests, selection of subjects based on their previous experience and performance in

> previous tests and elimination of subjects based on a statistical analysis of pre-tests."

In this study, an audiometric test is performed to verify that the users were normal-hearing and in particular, for older users, that they respected some characteristics useful for the test.

- **Post-screening of subjects:**

  > "Post-screening methods can be roughly separated into at least two classes; one is based on inconsistencies compared with the mean result and another relies on the ability of the subject to make correct identifications." In this case, researchers rejected subjects with valuations enormously different from the average results or with the inability to make correct identifications. It was done with the box-plot method to locate outliers.

- **Test method:**

  > "To conduct subjective assessments in the case of systems generating small impairments, it is necessary to select an appropriate method. The "double-blind triple-stimulus with hidden reference" method has been found to be especially sensitive, stable, and to permit accurate detection of small impairments. Therefore, it should be used for this kind of test."

  As already mentioned, this is the method used and explained in the following sections.

4.2 Selection of participants

In this research project two groups of subjects are tested:

- Elderly people with presbycusis, without otological pathologies, over 65 years of age;
- Young non-native speakers.

To make sure they can make good choices during the test, they are subjected to an audiometric pre-screening test conducted by the otolaryngologists of the University of Turin. This step is explained in detail in the next section. The number of elderly subjects is 41; 10 of these are part of the preliminary test pilot, used to choose the transfer function used in the main tests. As shown in section 5.2, the

"Heavy" t.f. was the most appreciated one. The young non-speakers tested are 12. Participation in the study was spread through the Doodle tool of Google, through which each person could book at a specific time slot.

4.3 Audiometric Test: liminar tonal audiometry

Liminar tonal audiometry aims to find the minimum auditory threshold for pure tones presented at liminar intensity[1] in the range of frequencies from 125 Hz to 8000 Hz. It is considered a subjective test because its execution requires the collaboration of the patient. During the audiometric performance examination in the soundproof room, the user is evaluated in the same sitting position that will be subsequently maintained during the performance of the subjective test, as shown in Fig. 4.1. The examination has always been performed by staff experienced in audiometric techniques (otolaryngologist or student in Medicine and Surgery adequately trained). The test is carried out using an audiometer, a pure

Figure 4.1: User subjected to audiometric test.

[1] *Liminar intensity:* minimum level of audibility of pure sounds perceptible by the human ear.

sound generator calibrated in "decibel hearing level" (*dB HL*) according to international standard (ISO 1975), supplied on loan free of charge by the manufacturer Inventis[16]. The sounds are administered to users, separately for each ear, by air via circumaural headphones and at the frequencies shown in Tab. 4.1. For each frequency the sound is sent at progressively decreasing intensity to obtain the last perceived intensity value, which corresponds to the limit of audibility for that given frequency. Figure 4.2 shows some examples of the control panel screens shown on the audiometer to the doctor in charge.

Frequencies of sounds administered to the user (Hz)							
125	250	500	1000	2000	4000	6000	8000

Table 4.1: Audiometry frequencies administered to the users.

(a) *Example of audiometric test result for left ear.*

(b) *Example of audiometric test result for right ear.*

(c) *Example of pure tone at 1 kHz at 25* dB HL.

Figure 4.2: Example of screens shown on the audiometer.

The obtained data are reported in graphic form on an audiogram using an international symbol system recommended by the *American Speech-Language-Hearing Association*[17] and adopted in ANSI S3 21.1978[18], in order to obtain the complete trace related to the right and left ear. The results of each user were compared to the UNI EN ISO 7029:2002[19] guideline regarding the statistical distribution of the audiometric threshold according to age and gender: **5 subjects were excluded** from this study, as the average of the values did not fall within the 25° and 75° percentile of the reference Gaussian.

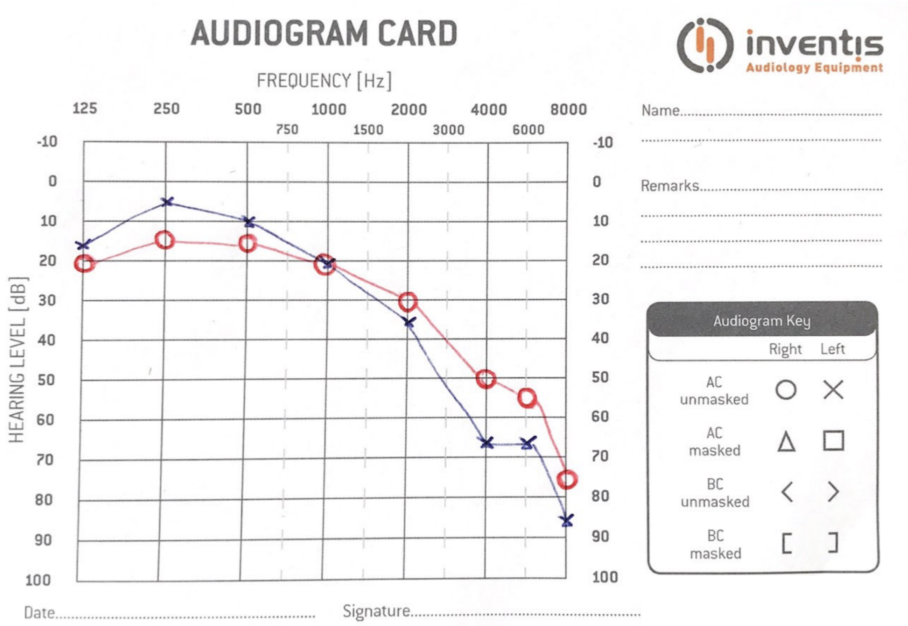


Figure 4.3: Example of audiometric test result. The figure shows if there are deficiencies from 125 Hz to 8000 Hz both for the left ear and the right ear. In particular, in this case the results show a typical presbycusis, with progressively increasing limit of audibility from 2 to 8 kHz.

Before the start of the test, each participant was provided with information about the objectives of the project and the main information regarding the conduct of the test. Furthermore, all participants were asked to sign the written consent to be submitted to the test and to be included in this study.

4.4 Listening test

This section describes the main listening test in which the user interacts directly with the extracts treated with the transfer function. The first part presents the methodology used to calculate the degree of appreciation of the improvement.

4.4.1 Test methodology and Subjective Difference Grade

The methodology used is the "double-blind triple-stimulus with hidden reference" recommended by ITU-R BS.1116-3[2] which says:

> In the preferred and most sensitive form of this method, one subject at a time is involved and the selection of one of three stimuli (“A”, “B”, “C”) is at the

> discretion of this subject. The known reference is always available as stimulus "A". The hidden reference and the object are simultaneously available but are "randomly" assigned to "B" and "C", depending on the trial. The subject is asked to assess the impairments on "B" compared to "A", and "C" compared to "A", according to the continuous five-grade impairment scale. One of the stimuli, "B" or "C", should be indiscernible from stimulus "A"; the other one may reveal impairments. Any perceived differences between the reference and the other stimuli must be interpreted as an impairment.

In the following sections, the stimulus "A" is called "Ref" (Reference). The "five-grade impairment scale" to which the standard refers are those shown in the Table 4.2 and refer to the Recommendation ITU-R BS.1284[20]. In this study, however,

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

Table 4.2: Five-grade impairment scale.

these grades fail to describe the user's judgment perfectly. It is crucial to understand if the processed audio track is perceived as better, equal, or even worse than the original reference track. For this reason a different evaluation system was used based on seven grades of judgment, which is always present in the recommendation BS.1284[20]. The analysis of user ratings is carried out through the

Impairment	Grade
Much better	3.0
Better	2.0
Slightly better	1.0
The same	0
Slightly worse	-1.0
Worse	-2.0
Much worse	-3.0

Table 4.3: Seven-grade impairment scale.

"Subjective Difference Grade" (SDG). As shown in the equation 4.1, it is based on the difference between the processed signal and the reference signal.

$$SDG = Evaluation_{Signal_{tf}} - Evaluation_{Hidden_{ref}} \tag{4.1}$$

In this case, the value +3 corresponds to 100% improvement in speech clarity; on the contrary, the value -3 corresponds to 100% deterioration. The SDG value is treated as a percentage value and calculated as shown in the equation 4.2.

$$SDG = SDG \cdot \frac{100}{3}\ [\%] \tag{4.2}$$

4.5 Training and Main tests

The training part (15 minutes) and the main test (30 minutes) are also held in the Audio Space Lab. The user placed in the same position as the audiometric test, 2 meters from the TV, has to listen to some excerpts from the RAI programs and evaluate the speech clarity (Figure 4.4).

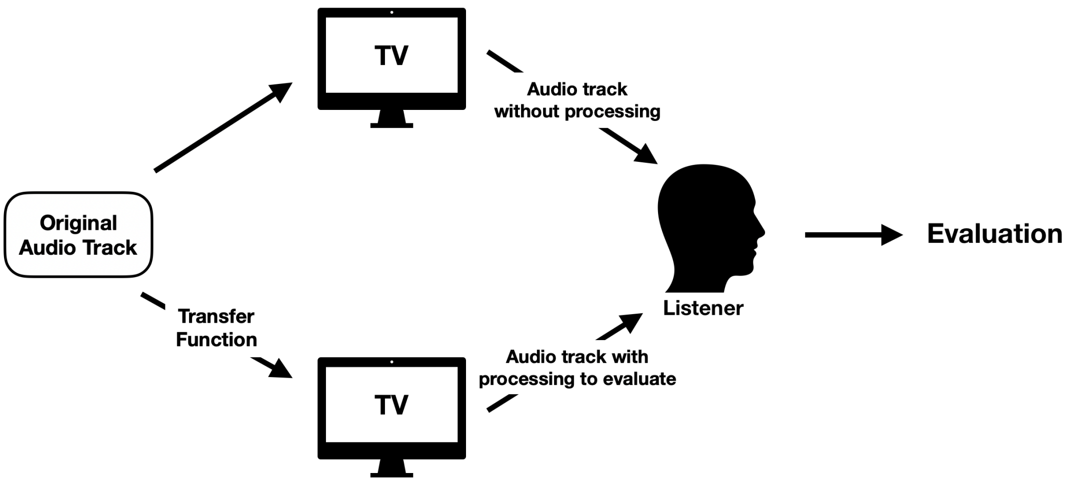


Figure 4.4: Diagram that represents the two signals (processed and original) that the user must listen and evaluate returning a value of percentage SDG.

In particular, it consist of:

- *Training phase (15 minutes)*: The test conductor explains how the test will take place. Three different tracks are presented, different from those present in the main test. It shows the interface with which the user can interact and make his considerations. At this stage, the test conductor stays in the room to clear all doubts and set the TV volume according to the user's needs. The volume should remain constant throughout the main test.

- *Main test (30 minutes)*: the user carries out the same evaluation procedure that he has tried during the training phase. In this case he have to evaluate 30 tracks and will be alone within the ASL, so that there are no subjective conditioning during the performance of the test.

Since the beginning of this treatment, the tracks have been understood as audio-video extracts. However, for this testing phase, it was chosen not to show the video part to users, but to play only the audio track of the different extracts. This is because having a visual part can mitigate the characteristics of the audio parts and influence the evaluation.
The interfaces (**Graphic User Interface** or **GUI**) the user interacts with, are two and have been created in C++:

1. *RaiTestTrain.exe* used in the training phase;

2. *RaiTest.exe* used in the main test.

These are explained in detail in the next two sections.

4.5.1 GUI for user data

The graphic user interface for user data is the first interface shown to the test conductor and is compiled after a series of questions asked to the user. Figure 4.5 shows the various fields, which are:

- *Test Number:* is an incremental number that identifies the test;

- *Age:* the age of the user;

- *Musical Experience:* the user is asked if he listens to a lot of music or if he has experience in playing musical instruments. This is because it can improve the ability to recognize improvements in audio tracks;

- *Watch TV?:* the answer can be yes or no;

- *How many hours per day?:* if the previous answer is yes, how many hours the TV is watched on average;

- *What TV programs?:* the category of programs that the user watches. For example, documentaries, films, news, etc.

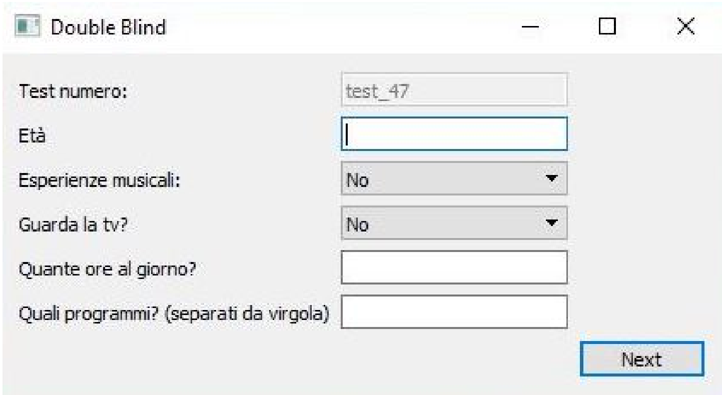


Figure 4.5: Graphic User Interface for user data.

4.5.2 Additional information

For statistical analysis purpose, other information was requested:

- *Gender*;
- *Years in Italy:* only for young non-native speakers;

The Delta_PTA parameter has been identified for the group of elderly subjects, namely the difference, for each subject, between the average reference Pure Tone Average, corrected for gender and age range, and its Pure Tone Average (PTA)[21]. The reference PTA has been calculated by averaging the minimum auditory threshold values on the frequencies 500 Hz - 1000 Hz - 2000 Hz - 4000 Hz, relative to the gender and age range of the subject, provided by the UNI EN ISO 7029:2002 guidelines[19]. The PTA of each subject was instead obtained by calculating the average of the minimum auditory threshold values of the two ears, on the same frequencies. A positive value indicates a lower auditory threshold than the standard PTA for gender and age range, therefore a better perception of pure tones. The components of the pilot, 6 males and 4 females, are aged between 65 and 78 years (average of 69.8 years old). The 31 subjects who participated in the final test are 22 males and 9 females, are between 65 and 85 years old (average of 72 years old). The Delta_PTA has a minimum value of -14.5 dB and a maximum value of +20.7 dB, with an average of Delta_PTA of all subjects equal to +1.7 dB. The non-native young people, 7 males and 5 females, are between 20 and 30 years old (average of 26 years old). The years of stay in Italy vary from 1 to 4, with an average of 2.5 years. The following Tab. 4.5 and Tab. 4.5 are the two tables used for the ANOVA statistical analysis, together with the SDG value results that will be accurately described in the next chapter.

User ID	SDG value (%)			Age	Gender	Delta_PTA
	Sport	Speech	Music			
1	0	0	0	66	M	8.05
2	-6	4	13	73	M	-8.05
3	38	37	27	70	M	7.58
4	47	33	33	77	M	7.58
5	-7	-12	33	70	M	20.71
6	28	21	20	74	M	1.96
7	7	10	0	85	M	-13.05
8	29	35	33	72	M	17.58
9	40	54	17	78	F	2.00
10	14	31	-33	65	M	1.80
11	33	43	89	72	F	7.00
12	28	27	17	77	M	7.58
13	-22	0	-25	65	F	2.05
14	-13	31	33	69	M	6.96
15	33	30	-67	68	M	-9.30
16	33	39	7	70	M	11.96
17	43	69	47	73	M	-14.30
18	-11	6	0	68	M	14.46
19	0	21	42	82	M	-11.80
20	11	46	50	77	F	-5.50
21	-5	-4	-8	64	F	-9.21
22	38	41	50	68	F	7.00
23	22	36	47	82	F	5.13
24	20	29	20	62	M	-14.46
25	8	39	22	67	M	-4.30
26	53	48	75	69	M	5.08

Table 4.4: Data of older users used for statistical analysis ANOVA.

4.5.3 GUI for the Main Test

After entering the user's data, the main test screen shown in Figure 4.6 appears.

User ID	SDG value (%)			Age	Gender	Delta_PTA
	Sport	Speech	Music			
1	33	44	-40	22	M	2
2	0	30	22	29	F	3
3	56	27	58	26	M	3
4	39	47	53	27	F	2
5	7	27	-20	32	M	2
6	5	13	-60	30	F	3
7	-10	39	17	24	M	3
8	28	10	40	20	M	1
9	-11	-18	-27	27	F	3
10	33	41	44	24	M	1
11	7	22	67	26	M	4
12	24	57	0	26	F	3

Table 4.5: Data of non-native speaker users used for statistical analysis ANOVA.

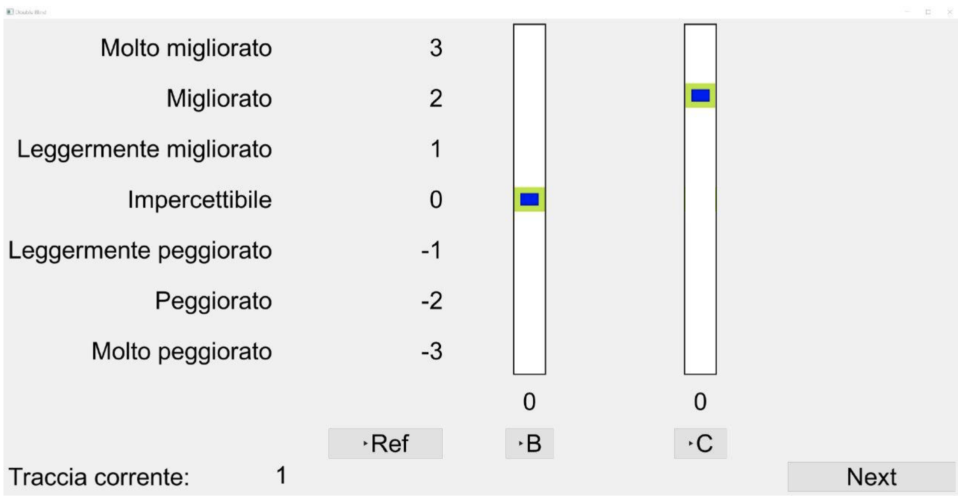


Figure 4.6: Graphic User Interface for the main test.

There are three buttons: "Ref" serves to reproduce the original reference signal without any processing; "B" and "C" are the two signals to be compared with the other. It is done according to the mode shown in section 4.4.1. Ratings can be made by dragging up or down the cursor relative to each signal, based on the seven degrees of satisfaction shown in Table 4.3. As shown in Figure 4.7, the user is provided with a mouse with which he can interact directly with the GUI. The "Ref" signal may be played up to 5 times to avoid an excessive duration of

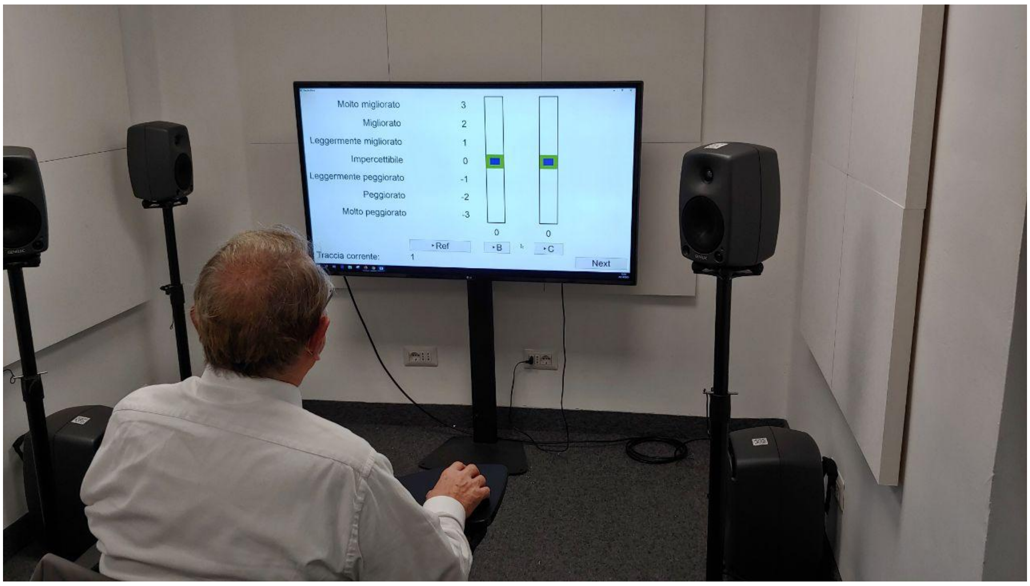


Figure 4.7: User during the main test.

the evaluation session. If exceeded, a pop-up window appears asking the user to perform the evaluation. Once the two signals are evaluated, the user can click on the "Next" button and proceed to the next track. At the bottom, in the middle, is the name of the extract of the program that is reproduced. The tracking number is visible at the left bottom, and it is progressively increased until it reaches the number of the thirty total tracks.
After the last track, the test program automatically generates a text file containing the ratings for signals "B" and "C", how much time was required to carry out the evaluations, and which track was the hidden reference.

Chapter 5

Results

5.1 Subjective Difference Grade calculation

The results obtained for each volunteer were obtained through a Matlab script that analyzes the text file generated by each test. The percentage of SDG is calculated using formula 4.2. Two text files are generated, for each of the two groups:

1. A text file containing the percentage of total average SDG for users;

2. A text file containing the percentage of total mean SDG for each of the 30 tracks analysed.

If the hidden reference has not been identified, or has been given a non-zero rating, the rating is automatically discarded from the total average. An example of SDG calculation is shown below.

5.1.1 Example for SDG evaluation

Table 5.1 shows an example of the possible combinations of SDG values detected by user votes. In particular:

- SDG_{test}: refers to the evaluation given to the signal subjected to the action of the filter

- $SDG_{hidden_{Ref}}$: refers to the valuation given to the hidden reference.

- $SDG_{difference}$: is the difference between the two previous values.

- $SDG_{percentage}$: is the value of SDG expressed as a percentage.

As mentioned above, if a valuation is given to the value of $SDG_{hidden_{Ref}}$, the value of SDG_{test} is not valid. The latter can take a positive value in case of improvement of the audio track treated with the transfer function or a negative value otherwise.

$\mathbf{SDG_{test}}$	$\mathbf{SDG_{hidden_{Ref}}}$	$\mathbf{SDG_{difference}}$	**percentage**	**Notes**
3	0	3	100%	The result is **valid** and represents the percentage improvement found after applying the filter.
0	2	-2	invalid	The result is **not valid** because it was given a value other than zero to the reference. The user is not able to recognize that the hidden reference is equal to the known one.
2	0	2	66%	The result is **valid** and represents the percentage improvement found after applying the filter.
2	-1	3	invalid	The result is **not valid** because it was given a value other than zero to the reference. The user is not able to recognize that the hidden reference is equal to the known one.
-1	0	-1	-33%	The result is **valid** and represents the percentage deterioration found after applying the filter.

Table 5.1: Possible voting results via SDG value.

This chapter reports the results of the SDG calculation obtained from the subjective tests, with also the statistical analysis ANOVA. Before this, preliminary test results are analyzed. They allow choosing the transfer function to be used for the two user groups to be tested.

5.2 Preliminary test

10 participants were divided into two groups during the first test day to choose which filter to prioritize: 5 participants with "Moderate" filter as reference, and 5 with "Heavy" filter. As described in the following two tables, the average SDG (%) per person was extracted. The results showed a higher SDG on average, meaning a more significant improvement in audio quality, with the "Heavy" filter (Tab. 5.3) than the "Moderate" filter (Tab. 5.2).

Participant	**Average SDG(%)**
$Test_1$	1.28
$Test_2$	4.44
$Test_3$	28.89
$Test_4$	15.67
$Test_5$	-3.45

Table 5.2: Preliminary "Moderate" filter evaluation test.

Participant	**Average SDG(%)**
$Test_6$	8.33
$Test_7$	-1.67
$Test_8$	34.92
$Test_9$	31.37
$Test_{10}$	8.77

Table 5.3: Preliminary "Heavy" filter evaluation test.

For this reason, during the following sessions, it was decided to test users with the application of the "Heavy" filter rather than with the "Moderate" filter.

5.3 Results per Track

In this section, the subjective test results for the two groups of users tested are exposed. The results are divided according to the 30 tracks used for listening sessions.

5.3.1 Elderly people

Figure 5.1 shows the average percentage SDG values per track for older users with "Heavy" filter application, orange for Speech, grey for Music, and blue for Sport. Furthermore, the desired improvement threshold of 20% is highlighted in red. Each media is also associated with the error bar indicating the standard deviation. As shown on the chart, there is an improvement on average for each track, with a threshold crossing for all tracks, except for track 2,3,4,6 for Sport, track 11 and 19 for Speech, and tracks 28 and 29 for Music. For the Music genre there is an average deterioration.

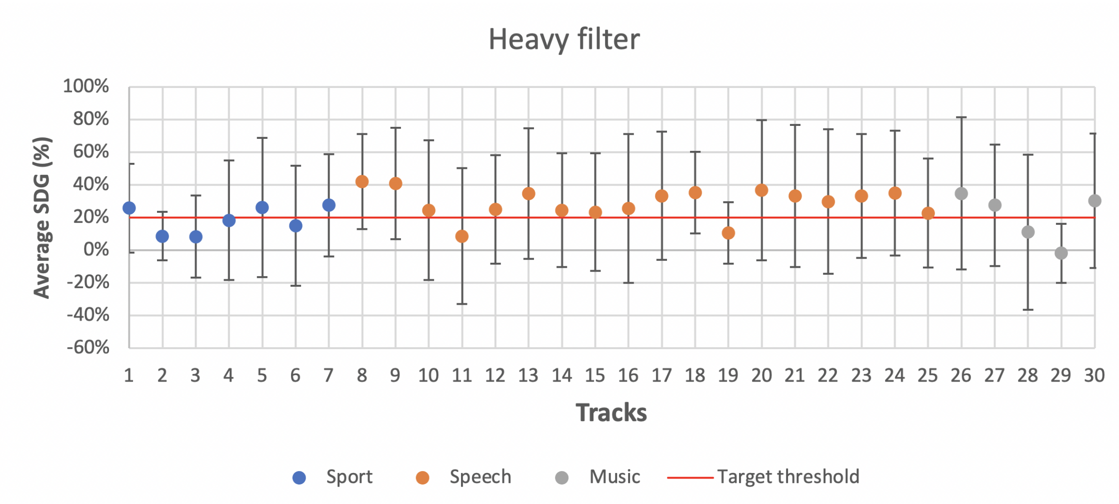


Figure 5.1: Average percentages SDG values per track related to older users with application of the Heavy filter and superimposed error bar. In red the threshold set to 20% improvement

5.3.2 Young non-native speakers

Figure 5.2 instead reports the average percentages SDG values per track for non-native young users. Also, in this case, there is an improvement on average for almost all tracks. Three of them worsened, one for every genre. The category in which the contribution of the filter is most appreciated is that of the Speech (in the next section, the chart is shown by genre), where some tracks have an improvement of about 50%.

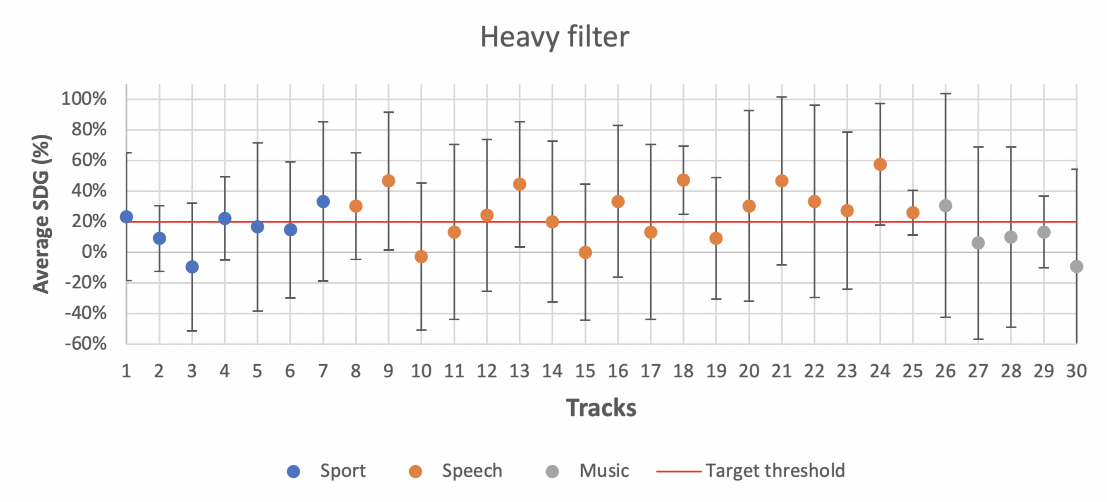


Figure 5.2: Average percentages SDG values per track related to young non-native speaker users with application of the Heavy filter and superimposed error bar. In red the threshold set to 20% improvement

5.4 Results per Genre

Through this section, the total average percentage SDG value divided by genre is evaluated.

5.4.1 Elderly people

The chart in Figure 5.3 shows the results related to the genre (Speech, Music, Sport), for elderly users with the "Heavy" filter. There is an improvement of 29% for Speech, 21% for Sport and 19% for Music. So, on average, they are all improved after applying the filter.

5.4.2 Young non-native speakers

Figure 5.4 shows the average results divided by genre for non-native young speakers using the "Heavy" filter. The results show an improvement of 28% for Speech, 16% for Sport and 10% for Music. Also, in this case, they are, on average, considered improved after the application of the filter.

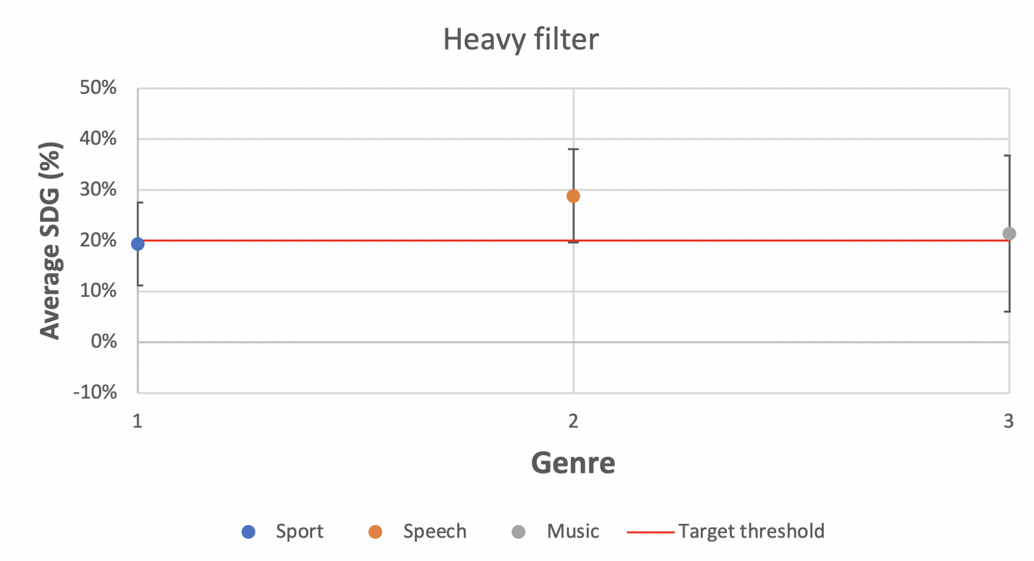


Figure 5.3: Average percentages SDG divided by genre related to older users with application of the Heavy filter and superimposed error bar. In red the threshold set to 20% improvement.

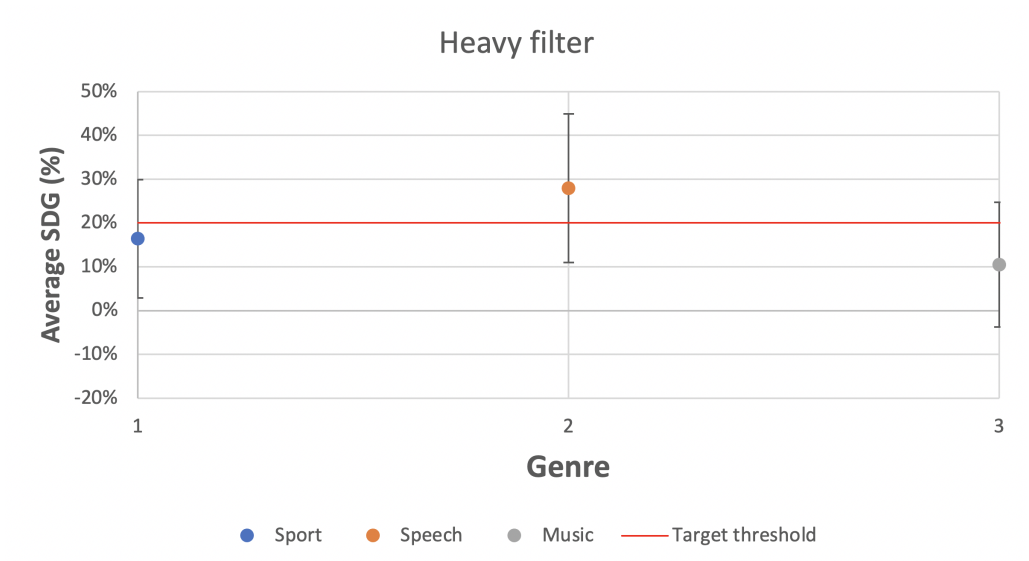


Figure 5.4: Average percentages SDG divided by genre related to young non-native speaker users with application of the Heavy filter and superimposed error bar. In red the threshold set to 20% improvement.

5.5 Statistical analysis - ANOVA

Once the evaluation of percentage SDG values was completed, the results were subjected to a variance analysis (ANOVA) using the SPSS statistical software. Data are normally distributed for both groups. Evaluations are made to identify significant differences between Genre (Music, Speech, Sport), Age, Delta_PTA for elderly users. A significance has been found in the factors Genre and Delta_PTA (Figure 5.5).
Regarding non-native young people, the groups Genre, Age, and Years of stay in Italy were analysed; no significance was found in these factors (Figure 5.6).

Tests of Between-Subjects Effects

Dependent Variable: SDG

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	48500,487[a]	74	655,412	6,022	,081	,993
Intercept	29943,289	1	29943,289	275,130	,000	,989
Genre	1466,520	2	733,260	6,737	,078	,818
Age	236,056	2	118,028	1,084	,442	,420
Delta_PTA	13295,722	10	1329,572	12,217	,032	,976
Genre * Age	574,444	4	143,611	1,320	,427	,638
Genre * Delta_PTA	10234,111	20	511,706	4,702	,114	,969
Age * Delta_PTA	,000	0	.	.	.	,000
Genre * Age * Delta_PTA	,000	0	.	.	.	,000
Error	326,500	3	108,833			
Total	86623,000	78				
Corrected Total	48826,987	77				

a. R Squared = ,993 (Adjusted R Squared = ,828)

Figure 5.5: ANOVA analysis for elderly users.

Tests of Between-Subjects Effects

Dependent Variable: SDG

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	27458,889[a]	32	858,090	,974	,606	,912
Intercept	9770,290	1	9770,290	11,086	,045	,787
Genre	726,118	2	363,059	,412	,695	,215
Age	11933,556	7	1704,794	1,934	,316	,819
Years in Italy	7251,500	3	2417,167	2,743	,215	,733
Genre * Age	6493,444	14	463,817	,526	,824	,711
Genre * Years in Italy	2557,333	6	426,222	,484	,794	,492
Age * Years in Italy	,000	0	.	.	.	,000
Genre * Age * Years in Italy	,000	0	.	.	.	,000
Error	2644,000	3	881,333			
Total	43870,000	36				
Corrected Total	30102,889	35				

a. R Squared = ,912 (Adjusted R Squared = -,025)

Figure 5.6: ANOVA analysis for non-native young users.

Chapter 6

Conclusions

The goal of the current project was to develop a transfer function (t.f.) to improve speech clarity on the most common flat-screen televisions available on the market. The project, funded by Rai, aims to be functional on its programs broadcast daily. So, for the tests, audio-visual tracks extracted from the television station's programs were used, divided into three genres: Speech, Sport, and Music.

The method refers to the ITU-R BS.1116-3 [2] standard, and is called "Double-blind triple-stimulus with hidden reference". The analysis was focused on the positive aspects of the "Heavy" t.f. (already implemented in the previous study) and its modified version "Moderate", on a group of elderly subjects with only presbycousis and on a group of young non-native speakers. At the end of the preliminary tests, the "Heavy" was, between the two functions, the one that provided the most positive results, and then it was chosen for the following test sessions. The difference between the two transfer functions is how they amplify some frequency bands. The "Heavy" amplifies the 4 kHz band more markedly, unlike the "Moderate" which amplifies that range in a more contained way and works more in the bands between 1.6 kHz and 2.5 kHz. The "Heavy" positive aspect is that various categories of population appreciate this transfer function, and the television broadcaster can only use this on programs that include different categories of users. In fact, in addition to the two groups tested in this study, the group of people tested in the previous project (normal-hearing subjects between 21 and 50 years of age) achieved a good Subjective Difference Grade (SDG) value, as explained below.

In this research project, the SDG results show an average improvement of 24.3% for the group of elderly subjects (standard deviation 10.9%) and 23.3% for non-native young speakers (standard deviation 17%) considering all 30 audio tracks analyzed. The results are very close to the SDG value obtained in the previous study, about 25%. The analysis of variance (ANOVA) through the statistical software SPSS has allowed to identify a significance in the factors Genre (Music, Speech, Sport) and Delta_PTA in the group of older people. The audio tracks related to speech showed a subjective improvement of about 30% for both groups of subjects tested. Through questions submitted at the end of the test, participants were asked in which categories they could better appreciate the contribution of the transfer function. Almost all users indicated the category of Speech, particularly TV fiction and film, and the Music category. However, although the modification of the tracks was more identifiable in the case of music, users complained about a degradation in the overall quality as the musicality of the song disappeared. Many users have also indicated the sports category in reference to sports news. In summary, the 20% threshold was therefore exceeded and the target reached.

Appendices

Appendix A

Inventis Triangle 11689-DD65 Screening Audiometer

- **CLASSIFICATION**: IEC 60645-1 / ANSI S3.6: Type 4
- **AVAILABLE SIGNALS**:
  - Stimulus: pure tone, warble tone
  - Masking (optional, with "Bone Connectivity" license): NBN, WN
- **SIGNALS SPECIFICATION**:
  - Attenuator step: 5 dB
  - Presentation: Continuous, Pulsed (1 and 2 Hz)
  - Warble: 5 Hz sine wave modulating signal
- **AVAILABLE OUTPUTS AND TRANSDUCERS**:
  - AC: DD45 or DD65 headphones, ER-3C insert earphones
  - BC: B71 bone vibrator
- **AVAILABLE TESTS**:
  - Manual Pure Tone audiometry
  - Automatic Pure Tone audiometry (available algorithms: HughsonWestlake, Quick Search, Fixed Intensity)
- **DISPLAY**:

- Type: Graphical color TFT LCD, capacitive touchscreen
- Size: diagonal 2.8”, 43.2 mm x 57.6 mm
- Resolution: 240 x 320 x RGB

- **POWER SUPPLY**:
  - Internal battery: Lithium-Ion rechargeable battery
  - Battery life: more than 12 hours of continuous use
  - Main Unit: 5 Vdc, 1.4 A
  - External dedicated USB adapter: 100-240 Vac, 50/60 Hz, 0.3-0.15 A, IEC 60601-1 compliant

- **PURE TONE: FREQUENCIES AND MAXIMUM LEVELS (*dB HL*)**

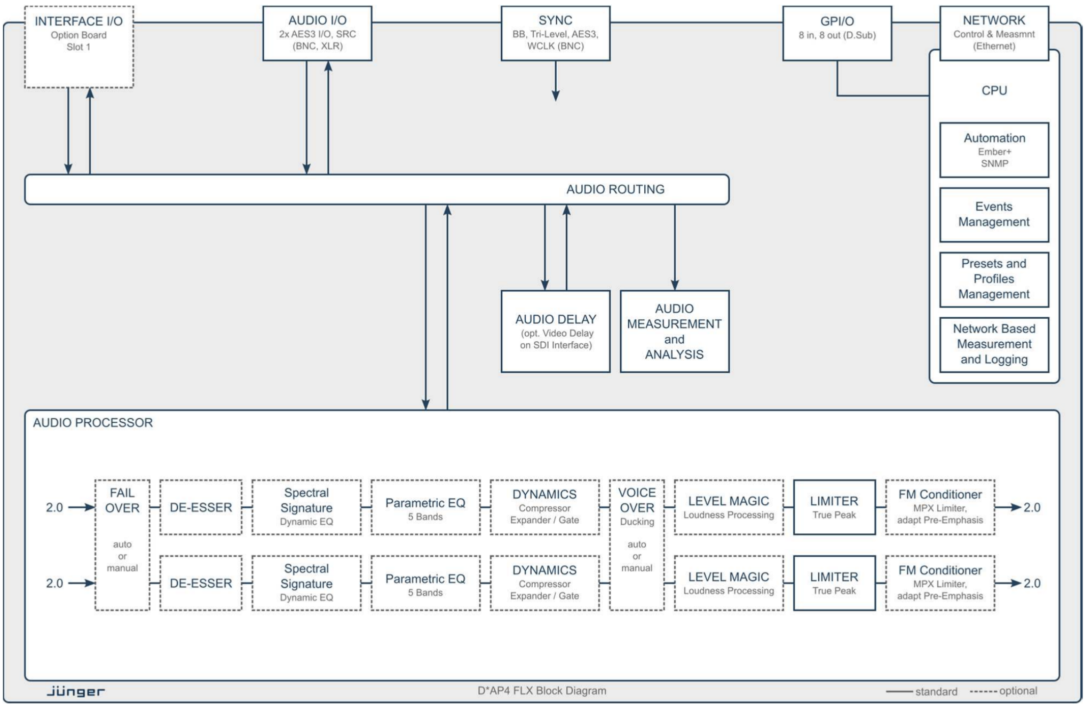
Frequency (Hz)	AC DD45	AC DD65	AC ER-3C	BC B71
125	65	65	80	-
250	85	80	90	35
500	100	95	100	50
750	100	95	100	55
1.000	100	95	100	60
1.500	100	95	100	60
2.000	100	95	100	60
3.000	100	95	100	60
4.000	100	95	100	60
6.000	95	80	90	40
8.000	85	80	75	35

Table A.1: Triangle Pure Tone, frequencies and maximum levels(dB).

Inventis website: `https://www.inventis.it/`

Appendix B

Junger D*AP4 FLX – 4 Channel Stereo Audio Processor



Level Magic Loudness Management: Based on a multi-loop control principle, the Level Magic algorithm provides adaptive wideband loudness control with

exceptionally high audio quality that is free of coloration, pumping, distortion or modulation effects.
Dynamics Processing: The dynamics section comprises of an upward compressor that controls dynamic range but retains the micro dynamic structure, and an expander/gate to remove low-level noise without introducing artifacts.
Parametric EQ and Spectral Signature: A 5-band Parametric EQ is coupled with Spectral Signature dynamic equalizer, a tool allowing adaptive control of spectral balance. Incoming audio is analyzed and its spectral structure is compared to a predetermined “reference” curve. This allows dynamic EQ corrections to be applied only if necessary to achieve a consistent sound image and tonal balance.
Fail Over: Fail Over ensures that if the primary signal is lost, the unit can switch automatically to a secondary backup feed.
Voice Over: Voice Over provides seamless integration of voice and program content in situations such as continuity announcements or video description with automatic ducking.
Loudness Measurement and Logging: The optional J*AM software displays input vs output levels and can log long term values to a destination folder anywhere on the network.
FM Conditioner: An option for FM broadcast to pre-control the MPX power and peak deviation of a stereo FM signal, including true peak and pre-emphasis limiting. A Phase Rotator automatically detects imbalanced audio waveforms and restores symmetry to prevent transients that may unnecessarily trigger downstream processing.
De-Esser: A De-Esser is available as a license option to control and manage excessive sibilance from spoken voices.
Web Configuration: A web interface allows setup and configuration from anywhere in the network. Onscreen metering and measurements are available for reference and a dedicated “mobile” UI is tailored specifically for touch screens or smaller displays such as tablets and smartphones.
System Integration: All system parameters are remotely accessible allowing the unit to be integrated and operated by broadcast control systems. A built-in Event Management tool allows the loading of parameter presets either by hotkeys on the optional X*AP RM1 remote panel, by 8 onboard GPI/O’s or by network commands using the Ember+TM control protocol. This helps users to apply individual processing to their programs or content and thereby ensure maximum audio performance in all situations.
Interfaces and System Security: Audio I/O’s range from onboard AES3 to optional 3G SDI including video delay, MADI, DanteTM audio over IP and analog. With power fail bypass relays, dual redundant PSU’s and SNMP integration, the unit ensures maximum operational safety and peace of mind for today’s critical

24/7 broadcast or content delivery facilities.

Bibliography

[1] Junger's manufacturer's website, 2021. URL https://junger-audio.com/en/products/slim-line/4-channel-stereo-audio-processor-d-ap4-flx.

[2] R ITU. Methods for the subjective assessment of small impairments in audio systems, 2015.

[3] Ebu r 128, 2020. URL https://tech.ebu.ch/docs/r/r128.pdf.

[4] M. Masoero A. Astolfi, A. Carullo L. Shtrepi, V. Cennamo F. Riente, and G. Barbero. Analisi e miglioramento della qualita audio percepita da tv a schermo piatto per la stima del subjective difference grade (sdg) a seguito dell'applicazione della f.d.t. *Politecnico di Torino*, 2019.

[5] Aja's manufacturer's website., 2021. URL https://www.aja.com/products/ha5-4k.

[6] Kanexpro's manufacturer's website., 2021. URL https://www.kanexpro.com/item/?id=SDI-SDHDXPRO.

[7] Andrea Prato. *Acustica dei piccoli ambienti*. Politecnico di Torino, 2019.

[8] Adobe audition website., 2021. URL https://www.adobe.com/it/products/audition.html.

[9] S3 22-1997 ANSI. Methods for calculation of the speech intelligibility index. *American National Standard Institute*, 1997.

[10] Herman JM Steeneken and Tammo Houtgast. Mutual dependence of the octave-band weights in predicting speech intelligibility. *Speech communication*, 28(2):109–123, 1999. URL http://www.sciencedirect.com/science/article/pii/S0167639399000072.

[11] CH International Electrotechnical Commission, Geneva. Sound system equipment–part 16: Objective rating of speech intelligibility by speech transmission index. *Standard IEC 60268-16*, 2020. URL `https://webstore.iec.ch/publication/26771`.

[12] S3 22-1997 ANSI. Methods for calculation of the speech intelligibility index. *American National Standard Institute*, 1997. URL `https://webstore.ansi.org/standards/asa/ansiasas31997r2017`.

[13] Stanley A Gelfand. *Hearing: An introduction to psychological and physiological acoustics.* CRC Press, 2017.

[14] Christopher A Brown and Sid P Bacon. Fundamental frequency and speech intelligibility in background noise. *Hearing research*, 266(1-2):52–59, 2010.

[15] Inail circular letter no.22 of 07 july 1994, 1994. URL `https://www.inail.it/cs/internet/docs/ci199422.pdf`.

[16] Inventis website., 2021. URL `https://www.inventis.it/`.

[17] American speech-language-hearing association (asha)., 2021. URL `https://www.asha.org/`.

[18] American National Standards Institute. Methods for manual pure-tone threshold audiometry, 1978.

[19] International Organization for Standardization. *Acoustics: Statistical Distribution of Hearing Thresholds as a Function of Age (ISO 7029: 2000).* ISO, 1991.

[20] BS Series. Recommendation itu-r bs. 1284-2. *ITU-R Recommendation BS*, 2019.

[21] Pure tone average (pta) definition, 2011. URL `https://www.nidcd.nih.gov/health/statistics/what-numbers-mean-epidemiological-perspective-hearing`.

Acknowledgements

Ringrazio la professoressa Astolfi per avermi dato l'opportunità di partecipare a questo progetto di tesi, insieme alla professoressa Shtrepi e al professor Masoero. Ringrazio il Dott. Riente per essere stato sempre disponibile a risolvere molti dei dubbi che incontravo durante il lavoro. Grazie al Dott. Andrea Albera e un grazie particolare a Caterina, eccellente collega con la quale ho avuto il piacere di collaborare.