POLITECNICO DI TORINO

Master's Degree in Mechatronic Engineering

Master's Degree Thesis

Gearbox fault and noise detection through audio analysis and deep learning

---

Supervisors:
Prof. Zampieri Nicolò
Prof. Somà Aurelio

Company tutor:
Ing. Bagatin Luca

Candidate:
Mariuzzo Claudio

A.Y. 2020/2021

Prima di iniziare a trattare questo elaborato, mi sento in dovere di dedicare questo spazio per ringraziare tutti coloro che mi hanno seguito ed accompagnato lungo questo fantastico percorso.

Innanzitutto, ringrazio l'azienda ROJ.srl, la quale mi ha dato l'opportunità di svolgere un tirocinio curriculare per la stesura di questa tesi. Ringrazio tutto lo staff, che è stato gentile e accogliente e si è dimostrato interessato al lavoro svolto.
Ringrazio il mio tutor, con il quale ho avuto la possibilità di mettermi alla prova e condurre le ricerche necessarie alla realizzazione dell'elaborato.

Un ringraziamento particolare invece, è rivolto a tutta la mia famiglia, a mia madre e mio padre, i quali hanno reso possibile questa esperienza e che mi hanno supportato in tutte le scelte prese in questi anni.
A mio fratello, che nonostante la lontananza mi è sempre stato vicino e si è reso disponibile nel momento del bisogno.

Inoltre, vorrei ringraziare 'Quelli di sempre' che da anni sono al mio fianco e mi hanno supportato anche nelle situazioni più estreme.

Infine, vorrei ringraziare tutti i miei amici, colleghi e compagni che mi hanno aiutato e sostenuto lungo questi cinque anni di università.
Vi ringrazio per tutti i momenti di spensieratezza, per la forza che mi avete dato e per i bellissimi momenti passati insieme.

Grazie davvero a tutti.

Abstract:

The achievement of high quality (Better-Best quality) and continuous research in minimizing waste (no Muda) are two of the fundamental points of the ideology of Lean production. ROJ.srl, careful to these considerations, has invested in this thesis for the research and design of an automated system able to test the quality of components purchased from external companies. In particular, the thesis focused on the defects and noise of the gearboxes purchased, aiming to have a system for the detection of gearboxes non-compliant, avoiding the use of these units in the assembly process and thus, improving the optimization of quality and waste of components at the same time.
The concept behind this project is to develop a physical system able to catalogue, through audio analysis, the gearboxes tested. For this purpose, the first chapters of the paper report a brief introduction to the reducers, the defects in consideration and hints of sound and signal theory. These topics are useful for the understanding and realization of the project. Subsequently, the design of the test bench was developed, with the plan, to install it during the acceptance phase. For the software implementation, two different approaches were considered: The first one, based on classical programming, has allowed cataloguing, with very high precision, the reducers noisy and defective results using innovative methods for the detection of periodic defects. While, in the second approach, a convolution neural network (CNN) has been trained, with various methods and strategies, for detecting defective reducers.
Also in this case, the obtained results were satisfactory, achieving a considerable reduction of the test time. Nevertheless, this solution led to a small loss in reliability, partially obviated, in a second moment, implementing a software based on the classification of several successive frames of the same tested component.
This last discussion has been very interesting in future development, as an onboard system could be implemented in such a way as to control, through predictive analysis, the occurrence of possible failures or defects.

Summary

List of Tables

List of Figures

1 Introduction

The proposed paper derives from the internship conducted at the Roj.srl company based in Biella. The main objective of the thesis concerns the study and the implementation of an automatic process for the analysis of gearboxes assembled on the DMD-0 brushless motor. This motor is controlled by an electronic drive and gets used in the field of agriculture applications, such as actuation for seed distribution in automatic seeding and fertilizer spreaders.

In the last production batches, a significant waste of material has occurred, compromising the normal production flow. The reducers, purchased by the supplier, reported some defects, so a non-negligible amount of motors was rejected at the end of the assembly line. To date, the quality control process, placed at the end of the line, is carried out by an operator, who, based on his own experience, judges the compliant motors. Nevertheless, this process is very subjective makes the process slowly and questionable. For this reason, the research of this thesis has focused on the development and analysis of a process aimed at avoiding production waste. In particular, the design of a test bench pointed at the control of the components purchased by the company will be reported. In order to replicate the system developed in the supplier company, a limited budget was assigned to the project. To respect this constrain, it was decided to use a system based solely on the audio signal, which has been implemented components easily available in the market. Instead, the software developed, completely implemented in python language, has been designed to acquire, process and classify the gearboxes, in this way only gearboxes that comply with the required parameters can be used in the assembly process.

1.1 Gearbox

DMD-0 motors mount an orthogonal gearbox that reduces the rotation speed from the input shaft to the output shaft through two different stages allowing a 90-degree rotation between the input and the output shaft.

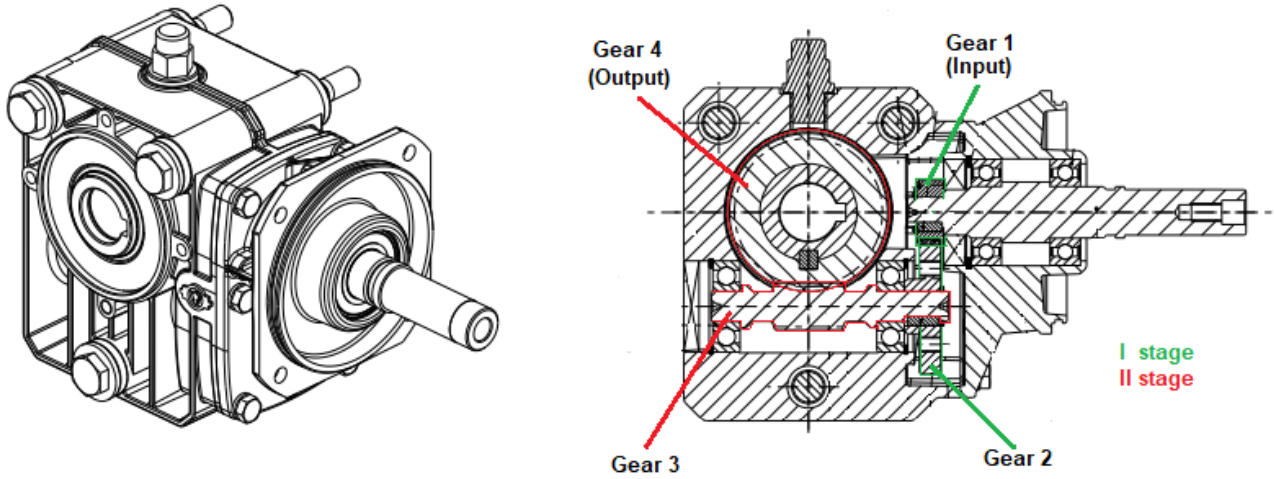


Figure 1.1.1: Gearbox configuration

The first stage is composed of a couple of helicoidal gears with Z1=48 and Z2=21 number of teeth, while the second one is comprised of coupling from worm shaft and worm shaft with respectively Z3=3 number of threads and Z4=39 number of teeth.

By knowing the number of teeth of the various gears, it is possible to obtain the transmission ratio of both stages such as a ratio of the number of drive wheel teeth to respect the number of driven wheel teeth:

Ratio of the first stage:
$$i_1 = \frac{Z_2}{Z_1} = \frac{48}{21} \tag{1.1}$$

Ratio of the second stage:
$$i_2 = \frac{Z_4}{Z_3} = \frac{39}{3} \tag{1.2}$$

Total ratio:
$$i_{TOT} = i_1\, i_2 = \frac{48}{21}\frac{39}{3} = 29{,}7114 \tag{1.3}$$

In this way, by imposing an input rotation ω_{IN}, we can discover all the rotation velocity of the gearbox components:

$$\omega_1 = \omega_{IN} \tag{1.4}$$

$$\omega_2 = \frac{\omega_{IN}}{i_1} = \omega_3 \tag{1.5}$$

$$\omega_4 = \frac{\omega_{IN}}{i_{TOT}} = \frac{\omega_3}{i_2} \tag{1.6}$$

This information will be very useful for the study and the analysis of the frequencies of the gearbox, as they will allow us to compute the fundamental frequencies of the reducer analyzed in *chapter 4.1*.

1.2 Problems encountered

During the assembly of the 'DMD 0 electric motor, in particular, during the testing and running-in phase, the operator reported loud noise coming from some gearboxes.

After closer examination more defects and sounds were found:

- The first taken into consideration consist of a periodic “click”, probably associated with a small surface imperfection on the teeth of the gears. In fact, we will see how this periodic sound is strictly associated with the rotation frequency of the second gear, (equal to third), and in some cases on the fourth one.
- The second detected problem, associated with a high level of noise, results a problem for the application. With the advancement of the analysis, It appeared that the increase of the noise level was caused, in some cases, by a concentration of high-frequency spectral components, while in the other cases the entire spectrum is more populated with respect to a normal gearbox.

The main goal of this project will be the development of an automatic system able to detect and report those types of problems.

At the beginning of the analysis, the assumption was that the ticking defect was caused by an error in the assembly phase of the encoder due to a high tightening torque that could damage one of the gears of the gearbox. To verify this theory was performed a static torque analysis aimed to identify the critical wheel causing the issue.

1.3 Static dimensioning

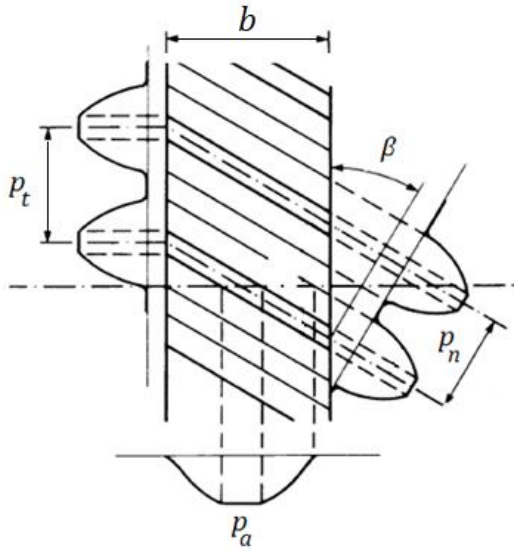
In the first phase of analysis, several motors judged defective were dismounted to be analysed and catalogued to understand the type the origin of the defect. The previously described problem, correlated with the ticking of the fourth gear, has resulted only in the dismounted components.
For verifying that wasn't a problem of production, static dimensioning related to the admissible input torque of the surface on the side of teeth was performed.
For this purpose, the geometric and mechanical characteristics of the reduction components are reported in the table below.

Gear	β [°]	β [rad]	α_n [°]	α_n [rad]	m_n	m_t	z*	z_v**	Y	b [mm]	λ	σr [N/mm²]
1	27,000	0,471	20,000	0,349	0,800	0,898	21	25,672	2,856	9	10,024	900
2	27,000	0,471	20,000	0,349	0,800	0,898	48	58,680	2,389	7	7,796	900
3	16,091	0,281	20,000	0,349	1,201	1,250	3	3,250	2,070	12	9,600	900
4	16,091	0,281	20,000	0,349	1,201	1,250	39	42,245	2,542	12	9,600	700

** In the case of gear 3, z represents the number of threads*
*** Virtual number of teeth was computed simplifying the formula because $\beta \leq 20°$*

Table 1.3-1: Mechanical characteristic of the Gears

In order to better understand this parameter, figure 1.3.1 is reported:



p_t = transverse pitch
p_n = normal pitch
p_a = axial pitch
β = helix angle of inclination
m_t= transverse module = $\frac{d_p}{z} = \frac{p_t}{\pi}$
m_n=normal module = $m_t \cos(\beta)$
z_v= number of virtual teeth =$\frac{z}{\cos^3 \beta}$
b = width band
$\lambda = b/m_t$
Y = Lewis number = $f(z)$
σ_r= tensile strenght [N/mm²]

Figure 1.3.1:Gear with helical teeth.

According to Lewis, the analysis carried out states that the minimum normal modulus of static resistance for helical wheels is equal to:

$$m_n = \sqrt[3]{\frac{2CY\cos\beta}{\lambda z_1 \sigma_{amm}}} \tag{1.3.1}$$

Since the dimensional characteristics are declared by the manufacturer in *table 1.3-1*, it is possible to derive the torque from *equation 1.3.1*.

$$C_{max} = \frac{m_n^3 \lambda z_1 \sigma_{amm}}{2Y\cos\beta} \tag{1.3.2}$$

The number of Lewis was obtained from the following table. Is important to notice that in the case of the worm gear, the number of the virtual teeth is equal to ∞.

Z, Zv	1 / Y	Y
12	0.245	4.08
13	0.261	3.83
14	0.276	3.62
15	0.289	3.46
16	0.295	3.39
17	0.302	3.31
18	0.308	3.25
19	0.314	3.18
20	0.320	3.13
21	0.327	3.06
22	0.330	3.03
23	0.333	3.01
24	0.336	2.98
25	0.346	2.89

Z, Zv	1 / Y	Y
28	0.352	2.84
29	0.355	2.82
30	0.358	2.79
34	0.371	2.70
38	0.383	2.61
43	0.396	2.53
45	0.399	2.51
50	0.408	2.45
60	0.421	2.38
75	0.434	2.30
100	0.446	2.24
150	0.459	2.18
300	0.471	2.12
∞	0.484	2.07

Table 1.3-2: Number of Lewis

Now, the torques obtained for each gear are reported to the input. In this way, it is possible to obtain the maximum applicable input torque, identifying the critical gear as the minimum input torque between all.
The results of the Lewis analysis are reported:

Gear	C [Nmm]	C_{in} [Nmm]
1	19056,25	19056,25
2	17719,42	7752,245
3	11288,36	4938,656
4	92941,84	3127,85

Table 1.3-3:Maximum admissible input torque results.

As expected, Lewis's results show as gear 4 is the critical component in this gearbox. In effect, this component is made of bronze-based alloy material, therefore is

characterized by lower mechanical properties than the surface-treated steel used for the other gears.

In this way, we were able to identify the tooth breaking limit corresponding to the maximum input torque applied. Nevertheless, for our purpose, it would be interesting to study the plastic deformation limit of the tooth flank. For this aim, a more conservative analysis, such as that of Hertz, has been applied.
Hertz's analysis is based on the Hertzian contact theory, which assumes the gear teeth as perfect cylinders of length equal to b.
The result of this hypothesis is that the static tension generated on the side of the meshing tooth is equal to:

$$\sigma_H^2 = \frac{1}{\pi}\frac{F}{b}\frac{R_1 + R_2}{R_1 R_2}\frac{E_1{}' E_2{}'}{E_1{}' + E_2{}'} \tag{1.3.3}$$

Where, E' is the modulus of elasticity with prevented lateral contraction, which takes into account the value of the Poisson's ratio ν. While, F is the total force of the contact, and R is the relative radius, which are calculated as:

$$E' = E/(1-\nu^2) \tag{1.3.4}$$
$$F = F_t \, cos\alpha_t \tag{1.3.5}$$
$$R = r \, cos\alpha_t \tag{1.3.6}$$

Substituting the above correlation in *equation 1.3.3* and Setting $\frac{1}{\pi}\frac{E_1{}' E_2{}'}{E_1{}' + E_2{}'}$ equal to K_e we can find the limiting tangent force of deformation as:

$$F_T = \frac{\sigma_H^2 \, b \, cos\alpha_t}{K_e} \, r_2 \sin\alpha_t \tag{1.3.7}$$

from the results obtained, it is possible to calculate the corresponding torque and the equivalent to the input.

Gear	E [MPa]	ν	HB	σ_H	Ft [N]	C [Nmm]	C_{in} [Nmm]
1	210000	0,3	578,5	1446,25	1584,614	14939,02	14939,02
2	210000	0,3	578,5	1446,25	2817,092	60704,58	26558,25
3	210000	0,3	654	1635	9574,031	71805,23	31414,79
4	118000	0,33	160	400	573,0326	13967,33	470,0543

Table 1.3-4: Hertz's analysis results.

The table shown confirms the results obtained previously, i.e., gear 4 is the critical component of the reducer. Furthermore, it is important to note that tooth flank deformation occurs for extremely low input torques. Despite this, this analysis indicates the first failures of the materials, with microscopic effects.

The result obtained from the Lewis and Hertz analyses proves how fundamental those verifications are in order to understand the origin of ticking problems. More precisely, during the encoder unscrewing, a tool is used to block the output gear that causes an

excessive resistive torque applied to the fourth gear. Probably, this screw tightness increase is due to the thread locker used during the assembly line, which would explain how the damage does not occur in the mounting phase of the encoder.

A practical experiment was carried out in order to confirm what was previously said. The test consists of the continuous increase of torque applied at the input, accompanied by the verification of the presence of the ticking sound. More precisely, by locking the output gear, a torque wrench was used to force the tightening of the encoder mounted on the input shaft. This test has allowed us to ascertain that for torques greater than 2.3 Nm the ticking occurs very clearly.

The result obtained broadly reflects that obtained from the analytical analysis. We must remember that the considerations made are the result of very important hypotheses and the result of the experimental test is indicative, as we cannot know to what level of deformation the presence of the tick corresponds.

2 Acoustic

The main objective of this thesis project is strictly correlated with the acoustic analysis, as it was decided to develop a system capable of categorizing the defects of electric motors through sound.
Acoustics is a branch of physics that studies the propagation of mechanical waves through a liquid or gaseous medium.

2.1 Sound definition

Sound is a mechanical phenomenon due to a vibrating behaviour of a source that causes the continuous compression and rarefactions of the air, through an elastic medium, around it. This continuous change in pressure in the region where they are acting causes longitudinal waves.

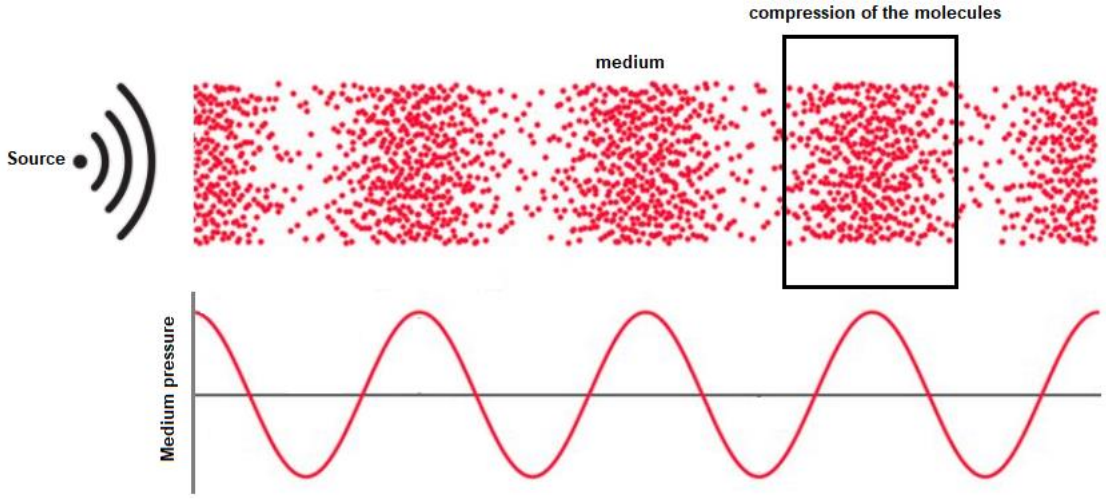


Figure 2.1.1: Mechanic of the sound

For humans, audible pressure is in the range of 10^{-6} $[Pa]$ to 10 $[Pa]$.
For fluids, the propagations of these waves are characterized by a finite value velocity, given by the Newton-Laplace equation:

$$c = \sqrt{\frac{K_s}{\rho}} \tag{2.1}$$

Where K_s is the coefficient of stiffness while ρ is the density of the medium.
In the case of conduction medium such as air in normal conditions[1], we obtain that $c = 343\ m/s$.

[1] *Since the air density and the stiffness depends on the condition of the air (temperature, humidity, pressure, etc) a standard reference as NTP-Normal Temperature and pressure is necessary. NTP is defined in air condition of 20° C and 1 atm with density of 1.204* kg/m^3. [1]

2.2 Waveform

The mechanical waves formed by a source are characterized by waveforms that propagate in all directions along straight lines, decreasing in intensity according to the inverse-square law. The waveform shown in *figure 2.1* represents a periodic-simple-single-sine wave (pure tone).

Sound, like any oscillatory phenomena, is characterized by a series of intrinsic characteristics, such as the frequency *f* equal to the inverse of the period *T* (in a periodic signal represents the minimum portion of time in which the signal returns to assume the same value), the amplitude *A* and the phase shift ϕ (translation of the starting point of the period).

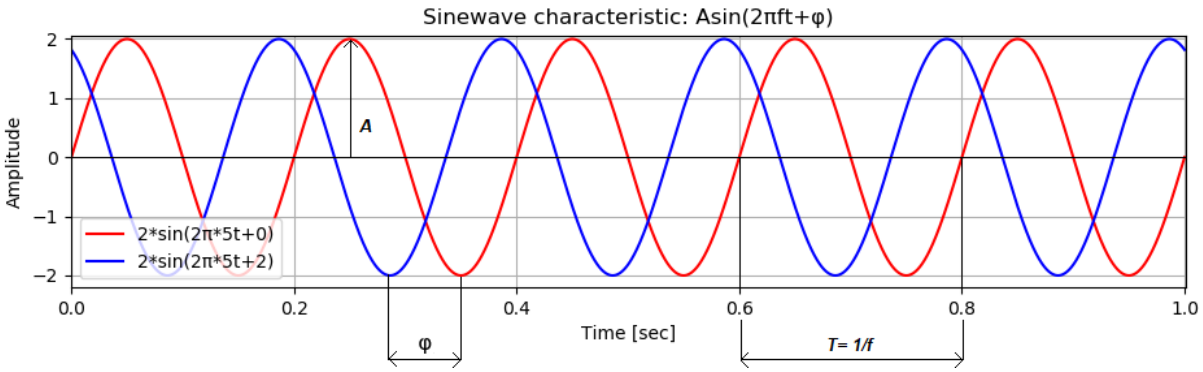


Figure 2.2.1: Sinewave representation

However, in most cases, the waveform representing a sound is more complex. Despite that, to obtain complex periodic waveforms it is possible to combine multiple sinewaves with different frequencies, phases, and amplitude.

In the case of no repetition patterns in the signal time-domain, another category of waveforms can be defined, the aperiodic waveform. This group of signals is composed of two main subclasses:

The continuous aperiodic waveform, also called noise, is characterized from a full frequency scale with random-similar values of amplitude.

A transient aperiodic waveform is defined by a brief duration with a higher density of frequency value similar to a 'click' [2].

2.3 Measuring sound

Above one of the fundamental properties of the sound was introduced, The frequency. Unfortunately, human perception of this characteristic is not linear, for this reason, the scaling for octave notation was introduced. It consists of a scale composed by eight octaves that are linear increases ranges of frequency. *Figure 2.3.1.a* shows how the frequencies are grouped. Moreover, it is possible to notice that each octave, wrt its previous, is formed by a double number of frequencies.

Nevertheless, the octave scale is not the only one. An example can be the PHON scale, which measures the loudness evaluated in SONES, or the pitch evaluation tanks to the MEL scale [3]. The first one introduced by American National Standards Institute is based on the concept of logarithmic measurements for the perceived sound pressure level of a 1-kHz pure tone that is judged as having the same loudness [4]. While the Mel scale was introduced to percept the pith of sound according to a series of filters in order to adapt the sound to the human ears, which finds many applications in the field of speech recognition.

Another type of measuring scale consists on measure the intensity, or amplitude, of waveforms. This technique takes the name of Decibel-scale. The term decibel, contained in the name of the scale, referred to a decibel that is a logarithmic unit of comparison, in this case, between two values of sounds pressures.

$$dB = 20 \log \frac{P_1}{P_2} \tag{2.2.1}$$

The reference pressure P_2 in air is set at the minimum capacity of human perception that is equal to $20\ \mu Pa$. In this way, it is possible to have a comparison meter of noises as the lower bound of this scale is equal to the minimum perceptible sound.

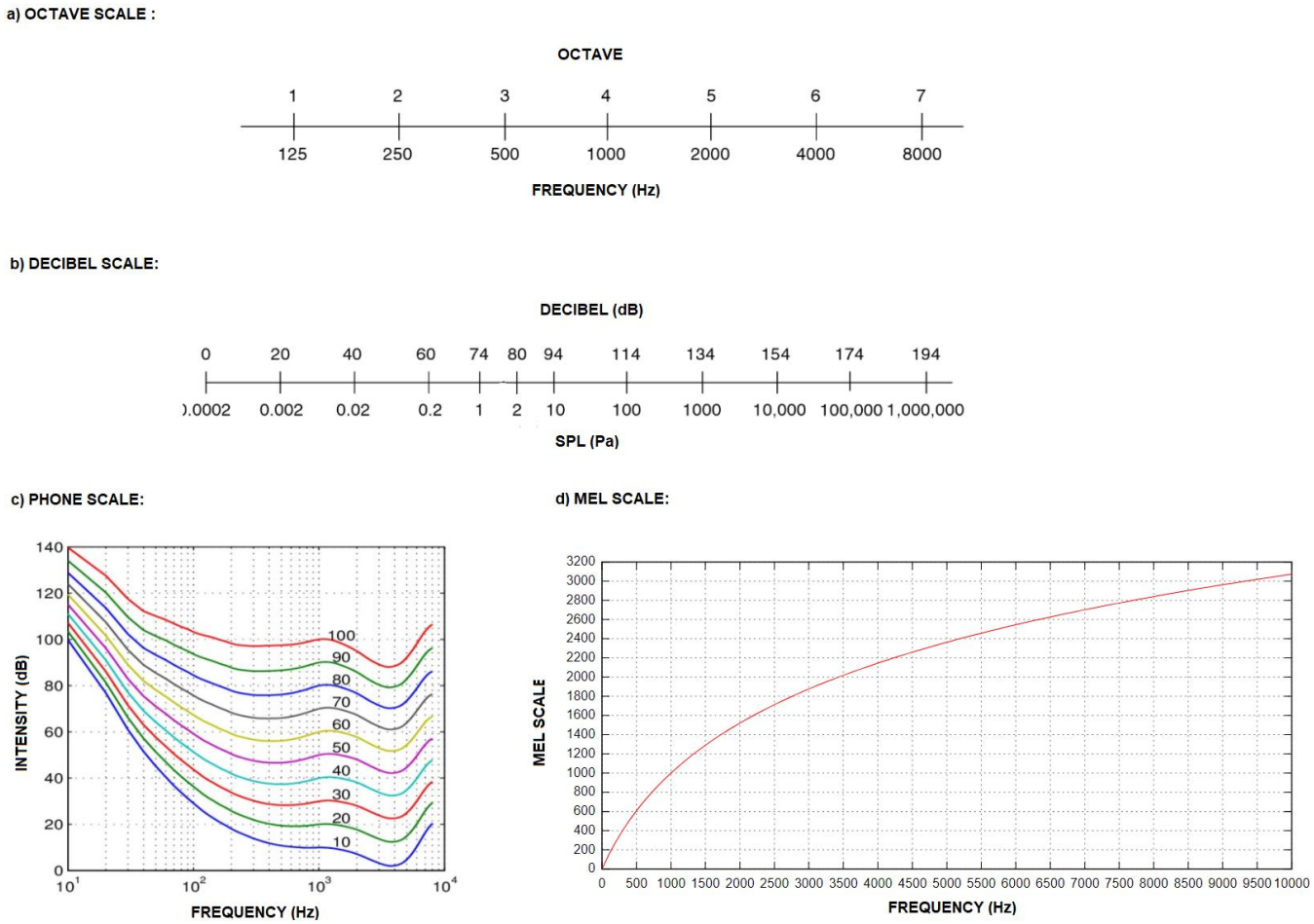


Figure 2.3.1: measuring scales: a) octave scaling of frequency levels. b) decibel scaling. c) phone scale. d) Mel scale.

2.4 Sound digitalization

Analog audio is represented by a continuous signal, with a theoretical infinite possible number of amplitudes and time. However, In the digital domain, a finite number of samples must be defined. Analogic to digital conversion provides to obtain a sequence of discrete values where data points are defined only on a finite number of values.

$$\underline{Analog\ domain} \quad \rightarrow \quad ADC \quad \rightarrow \quad \underline{Digital\ domain}$$
$$s(t) \quad \forall t \in R \qquad\qquad s(kt) \quad k = 1,2 \dots N$$

A/D conversion can be resumed briefly in two main steps represented from the sampling and quantization process of the audio waveform.

Sampling consists in to obtain, for each $t_n = nT$ point, a finite value of the signal, where $n \in Z$ is the number of samples, while T is the period of the sample.

In this way, it is possible to define the sampling rate S_r as:

$$S_r = \frac{1}{T} \tag{2.4.1}$$

It is easy to understand that a high value of sample rate corresponds to a better resolution of the signal with a consequent reduction of signal loss.

In order to get a proper measurement of data, an appropriate sampling time must be selected to avoid anomalies.

This decision should be based on the Nyquist-Shannon sampling theorem that is a theorem fundamental in the field of digitalization and data acquisition. It is based on the assumption of signals with finite bandwidth B, for this type of signal, a complete acquisition is performed only with a sample rate equal to the double of bandwidth $2B$, called Nyquist rate. From this theorem, it is possible to define the threshold frequency known as Nyquist frequency $f_s/2$ as:

$$2B < fs \rightarrow B < fs/2 \tag{2.4.2}$$

In the case of high bandwidth, a phenomenon called aliasing can manifest. In this situation, the implementation of a low-pass filter, also called anti-aliasing filter, can be fundamental to avoid distortions of the original signal.

The second step consists of quantization, i.e., the rounding-off of the amplitude of the sampled signal. As said for the frequencies, even for the signal amplitude value, the continuity cannot be satisfied. Therefore, a bit depth (number of bits representing the signal) is defined [5].
The results obtained by an A/D conversion are summarized in the figure below

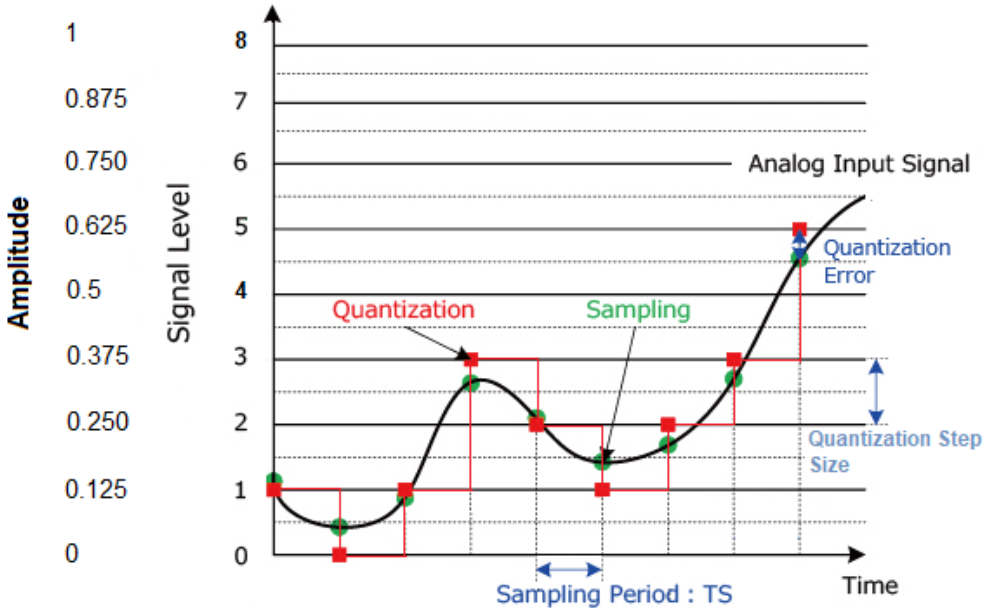


Figure 2.4.1: Quantized sampling with 8 representation levels (1 bit per sample).

Considering the above example, we have a continuous signal (black line) uniformly distributed between 0 and 1. The signal is quantized with $N = 3$ bit[2] so the difference between two adjacent steps is 0.125. The maximum error will be half of that. It is possible to confirm this intuition by applying the formula (2.3.1).

$$\mathrm{E}_q = \frac{\max(x) - \min(x)}{2^{N+1}} = \frac{1}{2^4} = 0.0625 \tag{2.4.3}$$

also in this case, by increasing the number of signal levels the resolution of the signal is higher and the quantization error decreases.

[2] *$N = 3$ bit correspond to the maximum representation value equal to $111 = 2^3 = 8$, in this case we have a signal x assume the value in the range of 0<x<1 then the amplitude step for each levels correspond to 0,125.*

2.5 Audio features

Audio features are the description of the sound, each different feature is able to provide a different aspect of sound. For this reason, those qualities are fundamental in the field of artificial intelligence (AI) and a complete study of them are very important in the analysis and categorization of audio.
In our case, in-depth knowledge of these properties will be useful for developing an automatic method able to recognize any possible defects presents on the gearboxes. In particular, a specific study of low and mid-level abstraction of sound will be considered.

The most important strategy of sound features categorisation is based on the signal domain that we are considering, as:

- Time-domain;
- Frequency domain;
- Time-Frequency domain.

These characteristics are introduced as they will be indispensable for the development of this project, moreover, good knowledge and acquisition of certain concepts will facilitate the understanding of strategies adopted for the analysis under consideration.

2.5.1 Time-domain features

Time-domain features are extracted from raw audio, which is represented from a waveform covered in chapter 2.1. Basically, we can take into consideration all the events that represent the sound. The time-domain features discussed in this chapter will be:

- Amplitude envelope (AE);
- Root-mean-square energy (RMS);
- Zero-Crossing rate (ZCR).

The first feature that we will look at is the amplitude envelope, which consists essentially of the maximum amplitude value of the samples present in every single frame.
We can define in this way the Amplitude envelope of a discrete signal $s(k)$ at frame t as:

$$AE_t = \max_{[k=tK\,;\,(t+1)K-1]} s(k) \qquad (2.5.1.1)$$

Where the amplitude computed at samples k is maximized between the first sample of frame t equal to tK and the last sample of frame t represented by $(t+1)K-1$.
This simple method can be applied to have a rough idea of the intensity of the signal that in our case is strictly correlated with the loudness of registration.

The second time features taken into consideration is called Root-mean-square energy which computes the RMS of all samples in a given frame and can be mathematically represented by the following equation:

$$RMS_t = \sqrt{\frac{1}{K}\sum_{k=tK}^{(t+1)K-1} s(k)^2} \qquad (2.5.1.2)$$

Also, in this case, the range of values is represented by the same set of values discussed before since at each instant the corresponding value of the sum of the square components of amplitude $s\,(k)^2$ is computed. The square amplitude can be associated with the instantaneous energy of the signal. So, the summation is the total sum of energy of the samples in the frame t, which divided by the number of samples on the frame become the mean of the sum of energy.
This feature is often used in signal analysis, in fact, like the first feature, that one can be associated as an indicator of loudness or can be useful in noise removal processes and segmentation.

The last, but not least, feature is the Zero crossing rate, which is more intuitive and characterizes the number of times where a signal crosses the horizontal axis. In other words, the ZCR is a function that allows to quantify the passage of the signal from positive amplitude values to negative amplitude values and vice versa. Also, in this case, it is possible to mathematically write this characteristic as follow:

$$ZCR_t = \frac{1}{2}\sum_{k=tK}^{(t+1)K-1} |sgn(s(k)\,) - sgn(s(k+1)\,)\,| \qquad (2.5.1.3)$$

Thank the equation above the last sentence is clearer. It is possible to notice that ZRC is computed as consecutive samples where a variation of the sign is detected, in fact in the case of two samples with the same sign the function returns a value of "0" meanwhile il the sign of the two consecutive samples are discordant the function returns "2" and this explains the presence of the multiplicative factor $1/2$.
this type of peculiarity is very useful in speech recognition processes, music processing and pitch estimation problems [6].

2.5.2 Frequency-domain features

As just said in the previous chapters, signals (with a special interest in audio-signal) in the time domain can be represented by the weighted sum of different sinusoids with differents phases and frequencies. For example, consider 3 different signals with respectively 0.2-0.5-1 Hz values of frequencies, when we sum them, we obtain a different signal.

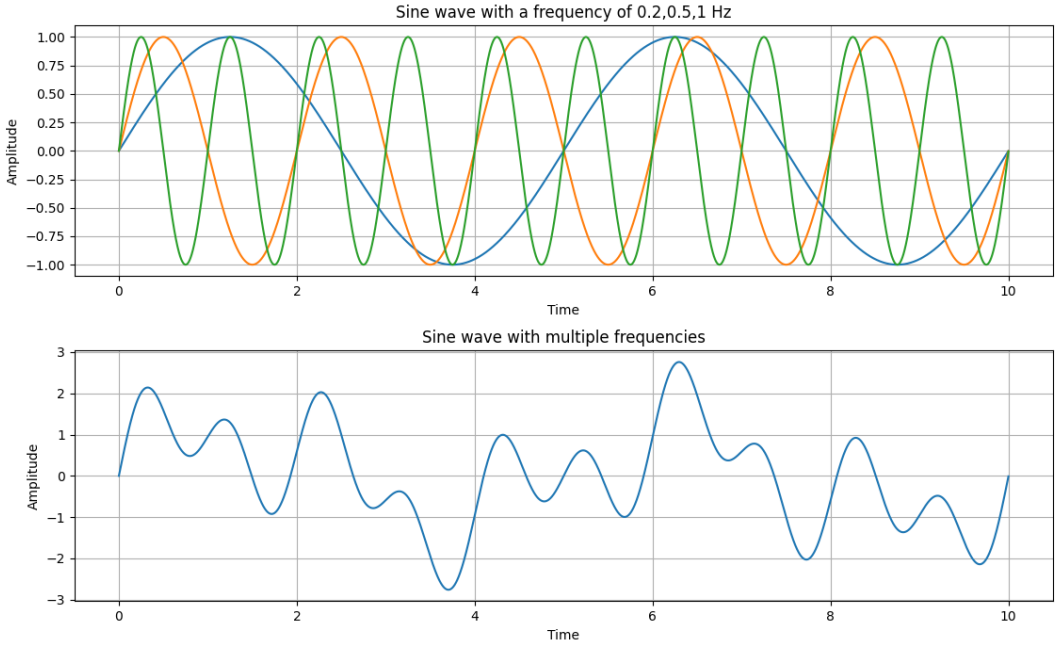


Figure 2.5.2.1: Signal as the composition of sines

The inverse process, or rather the decomposition in the fundamental sine waves starting from a signal in the time domain, of the above example, is represented by the Fourier equation.

Was Joseph Fourier in 1822, to formulate an operator able to represent any type of signal in the time domain into a frequency domain. In the case of a general continuous-time signal $s(t)$, integrable in the interval of $(-\infty, +\infty)$, thanks to the Fourier transform $\mathcal{F}[s(t)]$, it is possible to obtain the frequency contribution of the initial signal, according to:

$$\hat{s}(f) = \mathcal{F}[\mathrm{s(t)}] \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} s(t) e^{-i2\pi f t}\, dt \qquad (2.5.2.1)$$

This representation does not evolve over time, all the frequencies present in the signal are represented in a stable complex plot.

However, in the case of study of this thesis, the signal treated is not time continuous. Therefore, *equation 2.5.2.1* must be readjusted in such a way that it can be applied to a discrete signal.
In this regard, the discrete Fourier transform (DFFT) is introduced. Let's consider the result of audio digitalization discussed in *chapter 2.4*, where $x[n] = x(n\, t_n)$ with $n = 0,1,2 \ldots, N-1$.

$$\hat{x}(w_k) \stackrel{\text{def}}{=} \sum_{n=0}^{N-1} x[n] e^{-i\, w_k\, n\, t_s} \qquad (2.5.2.2)$$

The result of the above equation is an N equally spaced vector of the discrete frequency $w_k\ [rad/s]$, with a minimum step, represented by the resolution, equal to:

$$\Delta f = \frac{1}{N\, t_s} = \frac{f_s}{N} \qquad (2.5.2.3)$$

Appling the DFFT to the last example we obtain the following spectrum representing the frequency components of the signal.

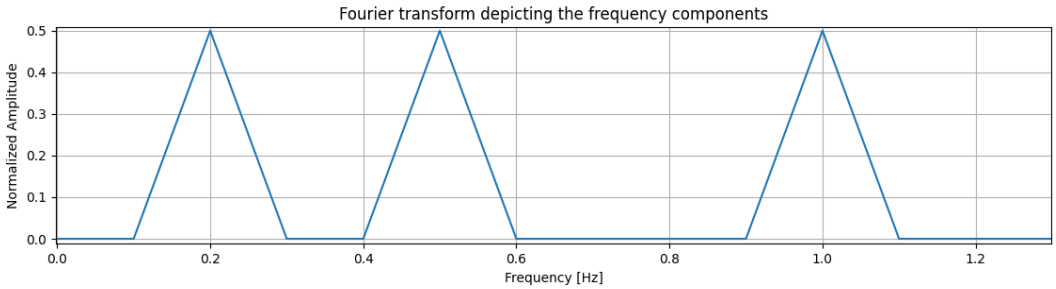


Figure 2.5.2.2:DFFT applied to the last example.

From the figure above, it is possible to notice three peaks in correspondence of the value of frequencies of the fundamental sinusoid of the signal. The amplitudes of the peaks are equal to each other because all contribution given from each sinusoid was equal.
Is important to take into consideration some limits of this theorem. In particular, using a discrete frequency, the time-limited of the measured signal must be taken into account. In fact, a sampled signal is rarely timed. For the acquisition, it is therefore essential to find a time frame in which to apply the FFT.
If the signal is interrupted, then in the case of non-integer numbers of periods to respect the signal, discontinuities occur which are translated into the frequency domain as high-frequency components. As these components are not really present in the original signal, the result is compromised.
Those discontinuities at the end of the signals, that are processed with the FFT, can cause the phenomena explained above, called spectral leakage.

One way to avoid this type of problem is to use a window in such a way as to modulate the signal at the end of the time window, in this way the contribution of the high frequencies will be much less. this technique is called windowing and consists simply in multiplying the acquired signal for a known function to smooth the ends points.
Given a discrete signal $S(k)$ and a discrete window $W(k)$ we obtain a windowing signal as:

$$S_w(K) = S(k)\, W(k)\ \forall\ k = 1 \ldots . N \qquad (2.5.2.4)$$

An example of windows can be represented by the Hann window, where thanks to the typical bell shape attenuate all values at the ends of the sampling period to zero.
A practical example will be reported when we will talk about signals acquisition in the case study.

2.5.3 Time-frequency domain

In this case, a combination of the last two domains is performed. An important advantage of this representation is due to the fact that is possible to have a temporal evolution of frequencies components of the signal. In the previous chapter, we talked about FFT which provide all information about the average frequency component across the duration of a specific signal segment, in this way is possible to know all the frequency components located in the signal, neglecting when are more or less present.
For this problem, a FFT is applied locally considering small segments of the signal called frames. This transformation takes the name of Short-Time Fourier Transform (STFT).
As we have seen previously, when we talk about signal framing it is good to take into account the windowing in such a way as not to report spectral leakage problems. In any case, an additional operation is required, that of overlapping, which allows not to lose any information during windowing, this happens thanks to the overlapping of the frames for a period sufficient to cover the ends of the frame where the signal was attenuated by windowing.

Remembering *equations 2.4.2.2* and *2.4.2.4* it is possible to formulate a mathematical representation of the STFT which is shown below:

$$S(m, w_k) \stackrel{\text{def}}{=} \sum_{n=0}^{N-1} x[n + mH]\, W(n)\, e^{-i\, w_k\, n\, t_s} \qquad (2.5.3.1)$$

It is possible to notice that unlike DFFT, which returns a series of complex coefficients as a function of the sampled frequency that represents phase and magnitude, while STFT depends on both frequency and time, where m is the number of time bins into which the signal is divided. Another significant difference is that in the case of *equation 2.4.3.1* the value of N denotes the number of the frame size and not the signal size as in the *equation 2.4.2.2.*
STFT returns a 2-D vector composed by the number of frequency bins equal to:

$$n°\ freq.bins = \frac{frame\ size}{2} + 1 \qquad (2.5.3.2)$$

and the number of frames represented by the following equation:

$$n°\ frames = \frac{samples - frame\ size}{hopsize} + 1 \qquad (2.5.3.3)$$

Of Complex Fourier coefficients.

Through this information, it is possible to plot and visualize the obtaining results the Fourier coefficients present for each item in the matrix are taken in square magnitude form and visualize it tanks to an indicative colour scale.
This graph is called a spectrogram and it will be very important for the development of the entire project. Therefore, good confidence with this graphic method will be indispensable for the understanding of the strategies adopted subsequently for the detection of defects.

For those reasons, a practical example is reported below in such a way as to have a clear pipeline for extracting the spectrogram from a signal and applying the technical issue encountered before.

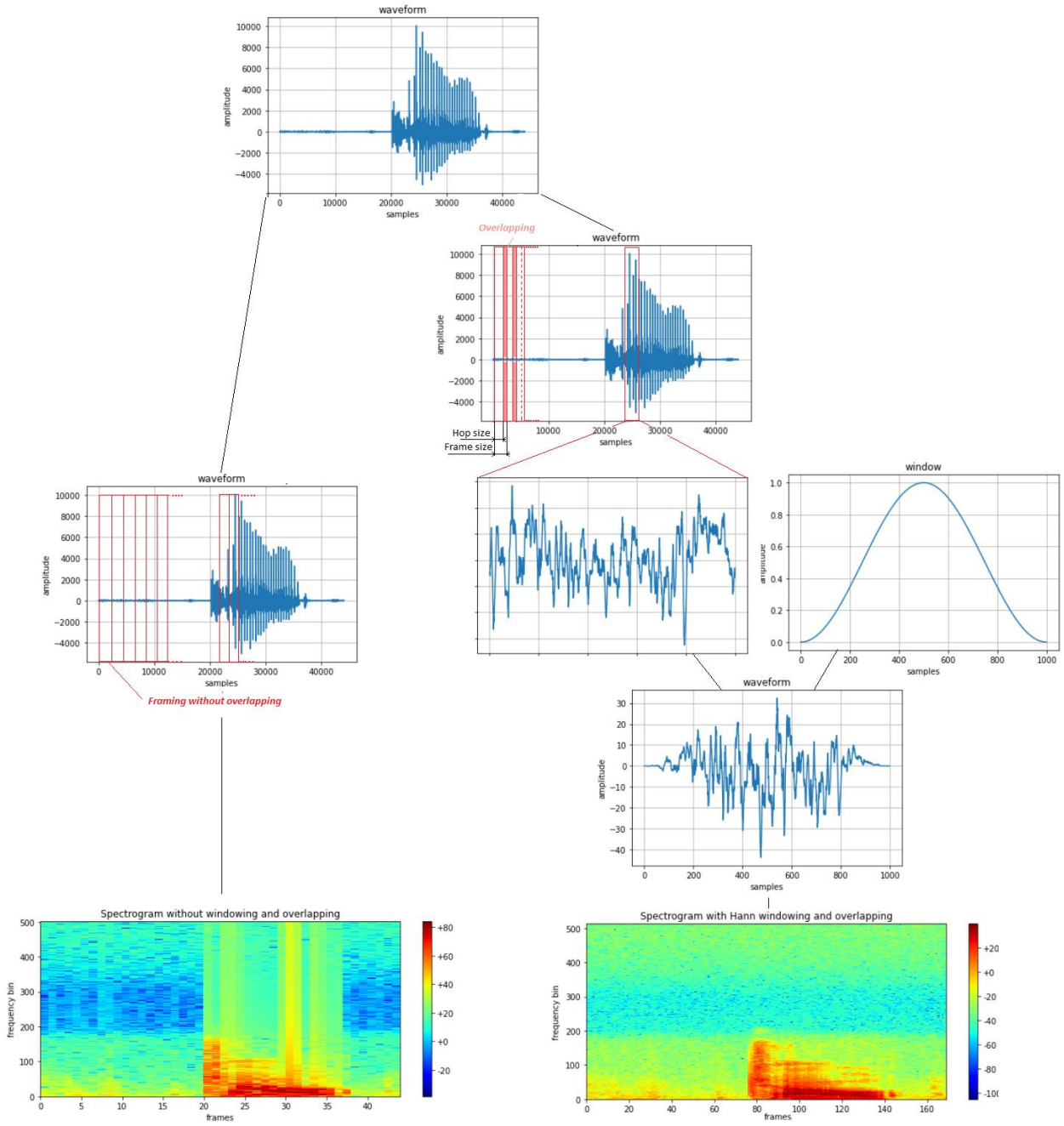


Figure 2.5.3.1: STFT application pipeline

The figure above shows two different approaches for obtaining STFT results.
The signal under consideration is audio of one-second length containing the "hello" word, the audio was sampled at 44100 Hz, therefore with a $t_s = \frac{1}{44100}\ s$.
On the left of *Figure 2.5.3.1*, a signal framing equal to 1000 samples without overlap has been applied. the result of this process is a matrix of 44 columns *(equation 2.5.3.2)* by 501 rows *(equation 2.5.3.3),* which represents the intensity of frequency, marked by the colour scale, over time.

In this case, it is possible to notice the phenomenon of frequency leakage due to the absence of windowing. while in the right part of *Figure 2.5.3.1*, it is possible to see how the obtained image turns out to be much sharper, thanks to the overlapping. The loss of information is much less, and the resolution turns out to be greater. In this case, we can apply the previous formulas and verify that with an overlap equal to 250 samples, the dimensions of the obtained matrix are: 501x173.

With the notions seen in the last chapters, it is possible to get to the heart of the project. the following chapter will begin to provide some information on the preliminary settings of the project followed successively by the chapters relating to the actual design of the test bench.

3 Audio acquisition

In this section, we will present a small introduction to microphones and audio capture in order to correctly choose the right device for the project. Furthermore, once chosen, the main characteristics and precautions for correct operation will be described.

3.1 Microphones

A microphone is a device symbolized by an electric transducer that converts the mechanical energy, represented by the sound pressure, into an electric signal. The main microphone that we will analyze are:

1. Dynamic
2. Ribbon
3. Capacitive
4. Piezoelectric

There are a lot of other types of microphones, but given the requirements on the cost and availability, those devices are not covered in this thesis.

Dynamic microphones:

The principle of this type of microphone is based on electromagnetic induction, more precisely the pressure air moves a thin component called the diaphragm. The diaphragm is attached to a coil wrapped up in a permanent magnet. In this way, according to the magnetic induction, a varying induced current is produced.
For this mechanism, this type of microphone is also called a moving-coil microphone.

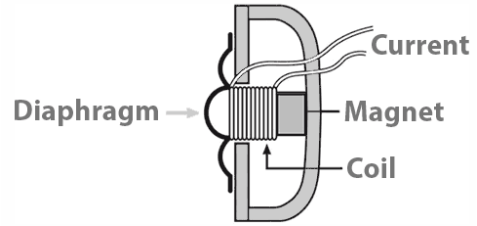


Figure 3.1.1: Configuration of a dynamic microphone.

Ribbon microphones:

A ribbon microphone is a type of dynamic microphone. The mechanism principle is the same, but he substitutes the diaphragm with a ribbon usually in aluminium suspended in a magnetic gap. The voltage generated is proportional to the speed of the ribbon. This configuration gives lightness, but at the same time fragility to the microphone. Furthermore, the cost of one of these microphones does not make it suitable for our application.

Capacitive microphones:

the capacitive microphone is based on a capacitor whose armatures are the diaphragm (mobile armature) and a rigid plate (fixed armature). the sound waves hitting the diaphragm move it by varying the distance between the 2 plates. Thanks to the relative

movement between the two, the capacitance of the capacitor changes and a different electrical signal are generated. The low mass of the diaphragm, compared to dynamic microphones, results in greater accuracy and sensitivity of the microphone. For this reason, the Capacitive microphones are the most popular in circulation, they are also available at a low cost.

Piezoelectric microphones:
This category exploits the response of ceramic materials to vibrations. The mechanic vibrations compress/expand the material that causes the production of the corresponding voltage. This configuration, based on the type of the material, is very sensitive to humidity and temperature, then the measurement can be distorted. Furthermore, the high impedance of the crystal makes the microphone very sensitive to noise. So, this compact configuration is an optimal solution for the acquisition of sound starting from the vibration of a body.

The first test bench prototype was designed with the aid of this type of microphone. The small size and low cost of this device seemed to make it great. despite this, after some tests, it was noticed that the microphone for optimal sound acquisition had to be attached to the reducer, more precisely on the back. The fixing was designed using a magnet, for its easy installation and removal. Unfortunately, the latter influenced the forces acting on the gear wheels of the gearbox, which produced an altered noise compared to the original one. For this reason, this path has been set aside in order to find a more optimal solution.

Directionality of a microphone
A microphone has the potential to pick up vibrations due to sound waves in the surrounding area, i.e., in all directions of 3-D space around the microphone. however, the construction characteristics of the microphones may have the purpose of favouring certain reception directions. This feature is fundamental for the categorization of microphones, in fact, depending on the use, one preferential direction will be better than the other.
This characteristic is described by the directional pattern, which is a polar diagram, where the direction associated with the intensity of the acquired signal is indicated. One aspect to take into account is that the polar pattern is a function of a given frequency (usually 1kHz) and becomes more and more directional as the frequency increases.

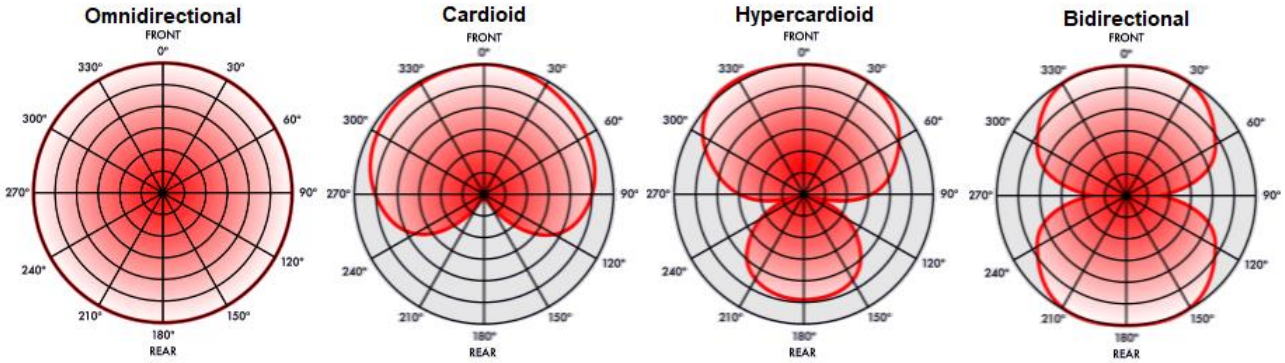


Figure 3.1.2: Polar patterns [7].

In the figure above the main polar patterns are reported, the first one represents omnidirectional mics which are equally sensitive in all directions. This type of microphone is not recommended for applications where there are disturbances or other noises external to the source of interest.
The second pattern is representative of a cardioid mic. Being much more sensitive in one direction (the front one), it is considered directional. The cardioid in fact has the characteristic of not capturing the sound waves coming from the back of the device.
Next, we find the hyper-cardioid which resembles the previous one, in that it has is predominantly directional along the frontal direction, but with a less developed lobe extending along the posterior direction. this indicates greater sensitivity for sounds coming from in front of the microphone. If both directions (front and rear) had the same sensitivity, then we would find ourselves in the last case or the presence of a bidirectional microphone.

Based on the information and characteristics described above, it was decided to purchase the MPM-1000U condenser microphone characterized by a cardioid type polar pattern, which allows, thanks to its positioning, to detect mainly the noise of the motor placed in front of the microphone, thus avoiding record noises from other parts. The frequency response between 20 and 17000 Hz widely covers the frequency band in question, with a fairly constant response. Furthermore, the USB input guarantees independence from the sound card characteristic of the computer used for data acquisition.
Other specific characteristics of the microphone are shown in the table below.

Figure 3.1.3: USB Microphone MPM-1000U.

Type	14mm back-electret condenser microphone
Polar Pattern	Cardioid
Frequency Response	20-17000 Hz
Sensitivity	-34 dB ± 2 dB (0 dB = 1 V/Pa @ 1 kHz)
Self-Noise	16 dBA
Maximum SPL	132 dB (THD ≤ 1%, 1 kHz)
Signal-to-Noise Ratio	78 dB
Power	USB

Table 3.1-1:MPM-1000U Technical specifications.

3.2 Acquisition method

For the acquisition of the audio signal, the microphone is located towards the rear of the gearbox, connected with the computer through a USB cable.
Since the signal on interest is not bounded, the audio is pre-processed to avoid error during the reconstruction of the signal. Moreover, the high frequencies characterized by a non-linear frequency response of the microphone are not taken into consideration. For those reasons, a low-pass filter is applied with a cutting frequency equal to 16 kHz.

In order to avoid aliasing, Shannon's theorem must be satisfied. Considering the bandwidth $B = 16\,\mathrm{kHz}$ of the signal, the sampling frequency must be greater than $2B$. In this specific case, we have decided to consider a sampling frequency of $2^{15}\ Hz$. In this manner, at each second 15 bits of samples are located.
The values of the samples are quantized with N=15 numbers of bits, this means that the values of the quantized signal can assume the values on the range of $\pm\, 2^{15}$. Remember the equation (2.4.1) the quantization error is equal to 0.5.

4 Test bench design

One of the arguments set for the thesis concerns the design of a test bench to test the gearboxes before assembly, i.e., in the acceptance phase.
The design involves the use of an electric motor, equipped with a drive, connected to the gearbox to be tested through an elastic joint. The system described above was subsequently placed in a soundproof box in order to avoid external sounds during audio acquisition.
The design of the test bench can be divided mainly into 2 distinct phases: the first, mainly mechanical, consists of the design of the locking supports for the pilot motor and the gearbox. The second part consists of the development of the software with the implementation of a program for driving the motor and acquiring the audio signal.

4.1 Mechanical design

The mechanical design of the test bench is focused on the ease of assembly of the gearboxes with a consequent saving of time for the operator.
Figure 4.1.1 shows the main components of the test bench. In particular, it is possible to recognize the three plates designed: the base, supported by 4 rubber feet to dampen vibrations, avoiding its propagation, and two supports facing each other.
The first support is equipped with a hole for the engine seat, supported with the help of 4 bolts.
The second has only a centring hole for positioning the reducer. In addition, an anti-rotation pin has been inserted to prevent the gearbox from rotating during torque transmission. For fixing it was decided to use a pair of clamps to facilitate the loading/unloading process of the gearbox to be tested.

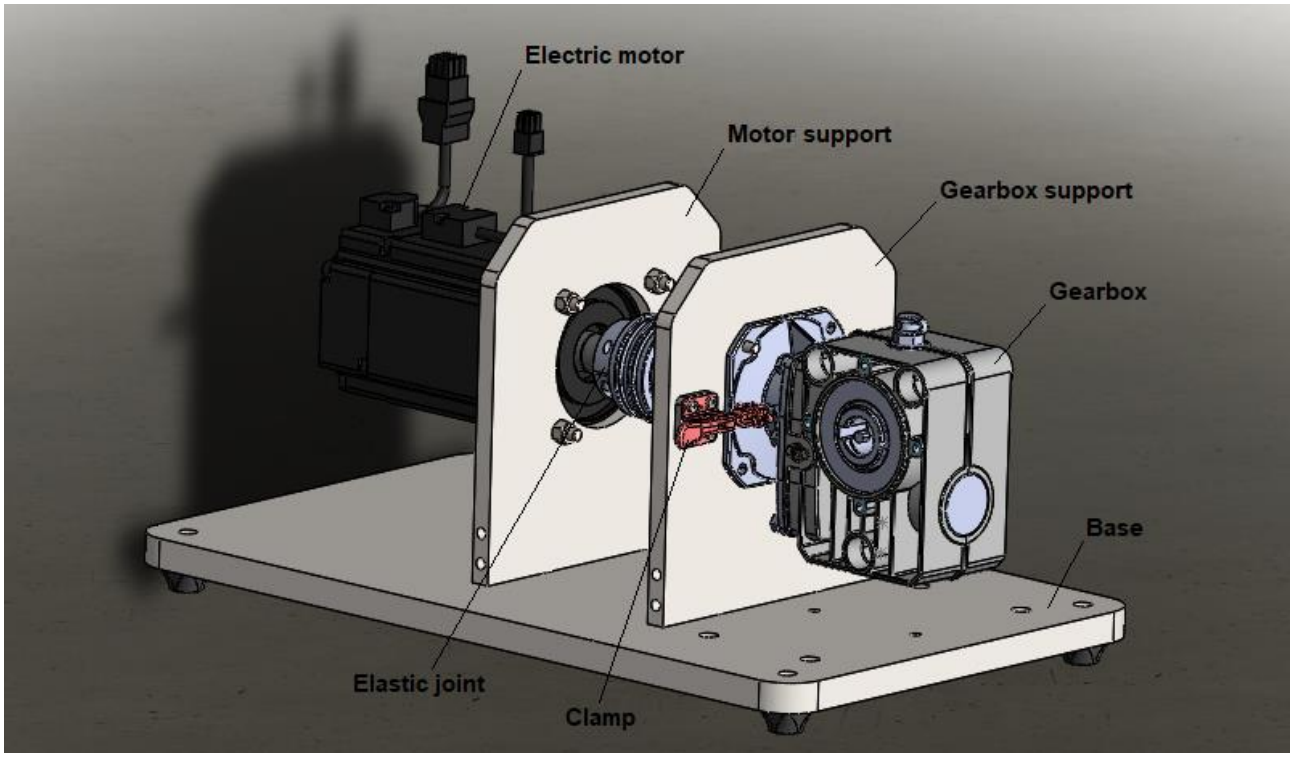


Figure 4.1.1: Assembly of the components related to the test bench obtained through Solidworks

Finally, the two supports were fixed to the base through a centring system composed of two pins and a central tap bolt. This choice is due to the need to have perfect parallelism between the supports, in fact, a minimum inclination of the two or an incorrect alignment of centring holes, present on the faces of the plates, would cause incorrect transmission, with an increase in noise and therefore with an alteration of the evidence.

Thanks to this configuration, in order to verify the conformity of the component, the test operator will only have to position the reducer, close the clamps and tighten the screw for locking the elastic coupling on the part of the reducer.

The choice of the servo motor was based on the availability of the components present in the company, with particular attention to the noise level generated. In fact, the use of a 'silent' motor is fundamental for this application, as being installed near the tested motor, could irreversibly alter the acquired data.
For these reasons the servo motor (ECMA-C80807RS), manufactured by Delta Electronics, was chosen.

The servo motor is driven by high performance with diverse communication interfaces servo drive ASDA-A2R series.The wiring method of the servo drive is based on a Single-phase power supply that supplies the main circuit (R, T) and the control circuit (L1c, L2c) with a voltage of 200-230 V.
The motor is connected in Three-phase through the U, V, W phases. The CN5 connector, connected to the motor encoder, allows to have a full-close loop or positional feedback connector depending on the chosen settings.
In the end, the CN3 connector ensures communication and motor control operating with Modbus communication protocol through RS-485 communication interfaces

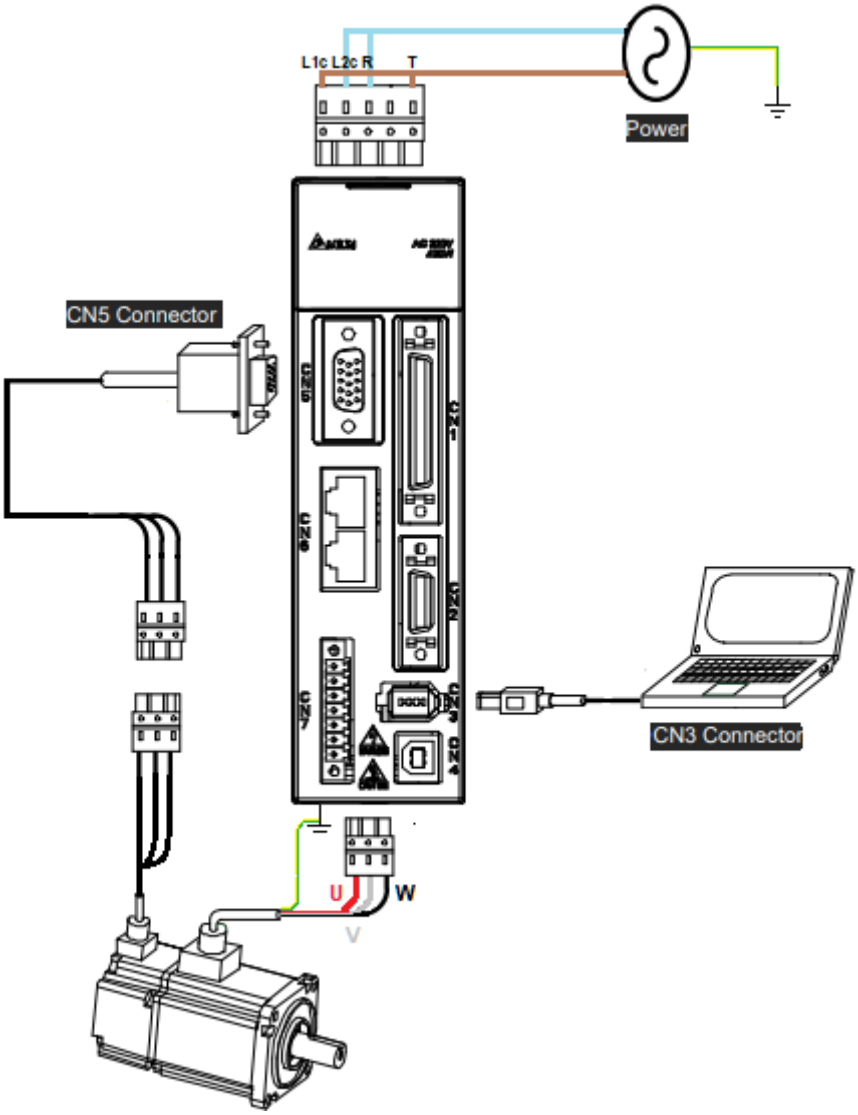


Figure 4.1.2: Wiring of Delta's Servo system with peripheral devices.

connected with the personal computer [8]. the communication of the engine can take place through the software supplied by the manufacturer, but for the project, the engine will be controlled manually through the python script described in the following section.

4.2 Software design

The entire software is based on the python programming language. This language, thanks to the vast libraries implemented, has made it possible to write a program capable of controlling the pilot motor and being able to connect the microphone for the acquisition and manipulation of the signal.

4.2.1 Motor control

As previously mentioned, the servo motor control was developed thanks to a python script. Python implements the Pymodbus library for the Modbus protocol. This protocol is today one of the most used protocols in the industrial field. The simplicity and reliability of this communication standard make it perfect for industrial applications.

Modbus is based on a type of Master/Slave communication, i.e., on a communication controlled by a device called master which controls one or more devices called slaves. These components are physically connected through two main different connections: the RS485 network, which uses the serial Modbus protocol and the Ethernet network, which implement the Modbus TCP/IP protocol.
In any case, the exchange of information is generated by the master, which, through a frame of bytes, sends the information or commands to the slaves through the field bus.

In our case, a Modbus protocol on an RS485 serial network has been adopted. Therefore, from now on, any reference to the Modbus protocol will refer to this category of a physical connection.

The complete description of this protocol can be defined according to the ISO/OSI standard model, which divides the communication protocol into 7 distinct levels:

Layer	ISO/OSI Model	Name of Protocol
7	Application	Modbus Application Protocol
6	Presentation	Not used
5	Session	Not used
4	Transport	Not used
3	Network	Not used
2	Data Link	Modbus Serial Line Protocol
1	Physical	EIA/TIA-458 standard

Table 4.2.1-1: ISO/OSI model of Modbus communication protocol.

As can be seen from the table above, in the serial Modbus protocol the intermediate layers are not used but are described in the case of an Ethernet network connection.

The physical layer is composed of a ring circuit that connects the master and the slaves connected to it in parallel. The transmission of information takes place through the exchange of 0/1 bits generated by a small potential difference present at the ends of the devices whose polarity defines the generated bit (-5V = 0 + 5V = 1).

The data link layer is responsible for the description of the transmission of the byte. The transmission can take place in two different ways: through a remote terminal unit (RTU) where the exchanged frames are composed of bytes with values between 0 and 255 with the need for timing for the synchronization of the frames, or with the aid of American standard Code for information interchange (ASCII) which is based on the hexadecimal encoding of each byte. In this case, the frame is characterized by a character that indicates the beginning and end of the frame, thus simplifying the management of the transmission but losing efficiency with respect to the RTU transmission protocol.
In our case, the RTU protocol was chosen as the transmission method. In this way, the master, in charge of starting the conversation, makes a request to the slave of interest, or periodically (polling) to constantly update the status of the registers.
The frame sent via the serial bus is made up of 11 bits which include a bit for the beginning of the frame '0', eight for the transmission of the communication byte, a bit for the control of transmission errors and a final bit with value '1'. In the event of a call from the master, the interrogated slave must answer within the pre-set time limit, otherwise, the master will pass to the next transmission.

The last level is represented by the Application layer, which deals with the transmission of data between master and slave called protocol data unit (PDU). This level differs from the previous one since the PDU contains the functions to be performed by the master on a slave. In fact, the bytes of a PDU is divided into several bytes in which the code of the function to be performed is shown in the first position, followed by the bytes containing the data useful for that specific operation [9].

for our purpose on python, the master was initialized by specifying that the transmission protocol was RTU, with a baud rate equal to 38400 b/s correspondings to the maximum speed supported by the RS-485 connection and response limit value equal to 2 s.
It was also necessary to set the motor parameters in such a way as to have direct control over the speed and the possibility to enable/disable as wished. To do this, various registers have been overwritten.
Finally, to avoid sudden acceleration/braking of the engine during speed changes, it was possible to modify the speed profiles in such a way as to define a trapezoidal speed profile as shown in the figure.

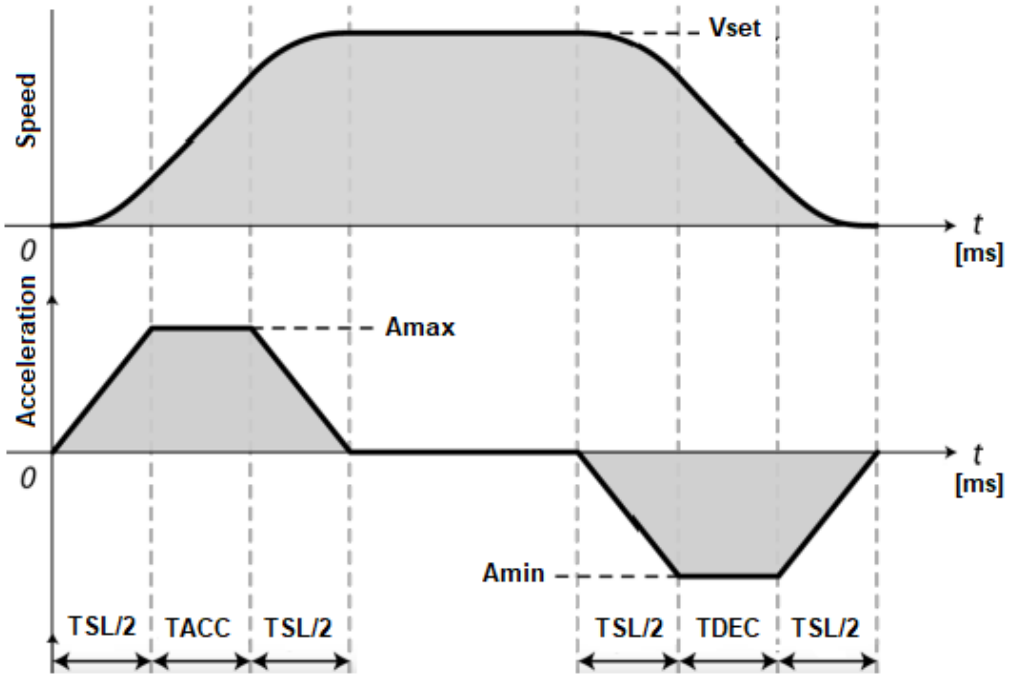


Figure 4.2.1.1:Acceleration/Deceleration Constant of S-curve.

With different parameters, it was possible to set the time dedicated to the various segments identified in *Figure 4.2.1.1.*
In this way the speed change is more progressive, avoiding high torque transmission over a short period which could damage both the electric motor and the gearbox tested.

4.2.2 Acquisition

For the acquisition of the audio signal, the python pyaudio library has been implemented, which allows to acquire audio by the microphone and save it in the standard digital format for audio, i.e., in the wav format (Waveform Audio File Format). The peculiarity of this type of format is the high audio resolution. Unlike other formats, the WAV format can save the sound without any compression, preventing the loss of information but weighing on the total weight of the file.
This solution made it possible to save all the registrations of the tested gearboxes, thus obtaining a dataset of useful information for the development and statistical calculation of the project. Furthermore, through recording, it was possible to listen in a second moment those audios to define the identifying characteristics of the various classes of gearboxes, without having to stop production and remove components useful for final assembly.

5 Test description

In this chapter, we will get to the heart of the project. In particular, the procedure used in the final software for the classification of the defective and/or noisy gearboxes will be exposed. Furthermore, some considerations and results obtained during the development phase of the project will be reported, referring to the topics dealt with in the previous chapters.

For the audio analysis, it was decided to perform a cycle consisting of 3 different speeds in both directions of rotation of the gearbox (clockwise and anticlockwise). more precisely, the speeds taken into consideration are 900,2100,3000 RPM. These speeds have been chosen in such a way that they do not have close correlations between them to avoid, during the design phase, situations that are difficult to understand, as phenomena of ghost spectrums or resonances. In addition, the maximum speed chosen corresponds to the highest travel speed for the application in which the gearbox is the protagonist.

5.1 Test results

At first, the analysis of the reducers has particularly focused on the study of the features in the time and the frequency domains. This process was helpful to evaluate different aspects of the acoustic analysis, placing the foundations for different approaches to the analysis of defective gearboxes.
In order to have an idea of how the test was carried out and what is the result obtained from the recording, the waveform of the entire signal has been reported in the following figure.

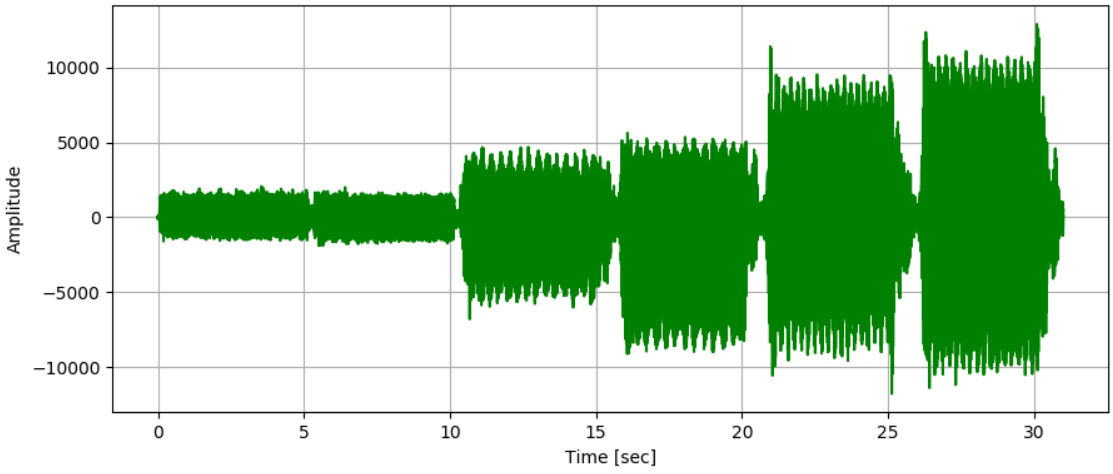


Figure 5.1.1:Waveform amplitude of the signal obtained by the test.

from F*igure 5.1.1* it is easy to see the speed changes mentioned above. Moreover, it can be seen how the behaviour of one direction of rotation faithfully reflects the opposite direction of rotation. The test was designed to last 30 seconds with the possibility of obtaining a wide range of data at different speeds. Despite this, the test lasts a little longer than expected. This is due to the set speed profile, as deceleration and

acceleration occur gradually. This does not appear to be a problem, but it will be necessary to pay attention in the frequency analysis, where the speed changes will be eliminated, taking into consideration only the four seconds where the speeds are constant so as not to consider any frequencies not proper to the operating speed.
Although the previous graph makes us understand the trend of the waveform over time, the information concerning the amplitude of the signal are not clear. For this reason, *Figure 5.1.2* shows the same signal but enveloping the amplitude and converting it to decibels.

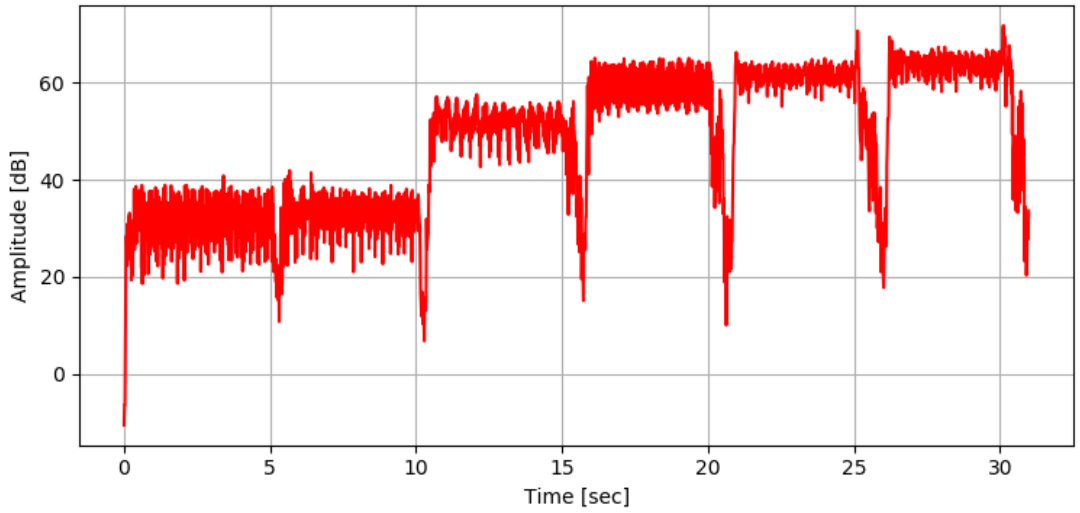


Figure 5.1.2: Waveform of the test signal in decibel.

the result reported was obtained by dividing the signal into frames of length equal to 512 samples and evaluating the maximum value in decibel of each frame.
Thanks to this representation it is possible to understand how the noise and therefore the sound pressure level increase as the speed of the gearbox increases. Furthermore, using this scale, we can have an index of how noisy the entire system is. In fact, in our case, for a speed of 3000 rpm, we exceed the threshold of 65 dB which is equivalent to the noise generated by a vacuum cleaner placed at 3 meters.

Another fundamental analysis for understanding the noise generated by the gearboxes is the frequency analysis. Specifically, the FFT was applied to the audio signal in order to obtain the frequency spectra relating to the various rotation speeds. To do this, as mentioned above, the recorded signal has been divided into 4-second segments equivalent to the operating speed. In this way, the extremes in which the engine is in the acceleration and deceleration phases are not considered.
In the figure below are reported the obtained results comparing the same velocities with the opposite rotation direction. In red are represented the velocities with a clockwise rotation, meanwhile in blue the counterclockwise direction.

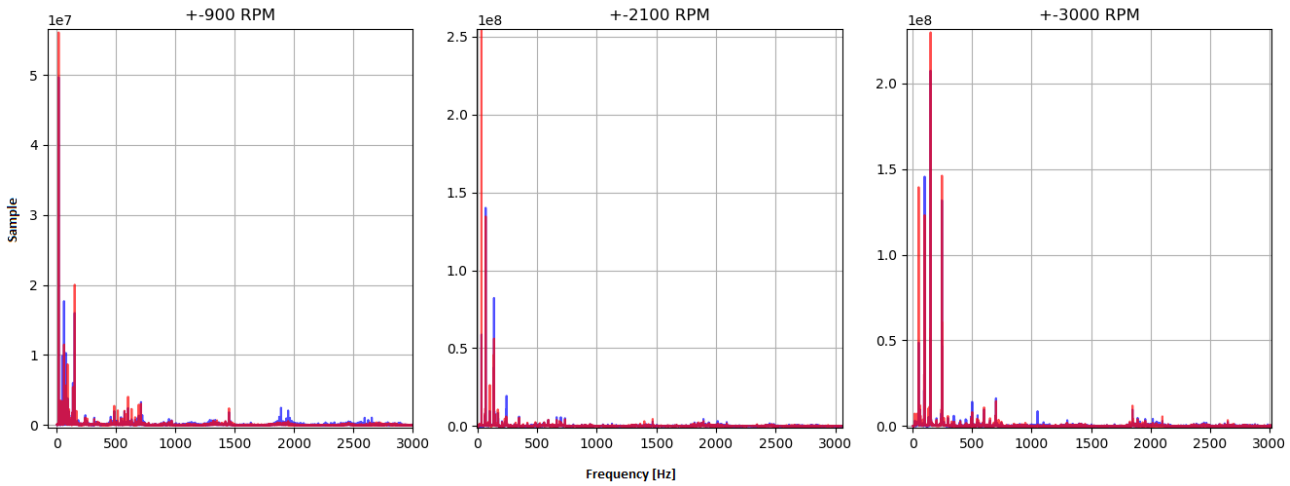


Figure 5.1.3: Frequency comparison between the rotation direction of the reducer.

As can be seen from the graph, the frequencies obtained for a speed are the same in both directions of rotation, with more or less comparable amplitude values. This is the confirmation that the nature of the noise generated by the reducer is the same in both directions of rotation.
To better understand the results obtained in the frequency domain, it is good to introduce some basic notions regarding the intrinsic frequencies of the gearboxes, i.e., the spectral components produced by the gears meshing.
The first frequency that we will consider is precisely the gears meshing frequency (GMF), which is the frequency of rotation of the gear proportional to the number of teeth of the gear itself. in a nutshell, this quantity defines the contact frequency of the teeth between a pair of gears and is obtained as:

$$GMF_i = f_i z_i \tag{5.1.1}$$

Where f_i represent the frequency in Hz of the gear rotation speed, where 1 rpm corresponds to $\frac{1}{60} Hz$. Naturally, the GMF value of the two coupled gears will be the same, since the repetition of the contact of the teeth must occur simultaneously.

There are other 2 frequency characteristics important to mention, but unlike the previous one, they are not always present and sometimes they can be smaller than they are negligible to respect the other components. these components occur at fractional frequencies concerning the meshing frequency. The first phenomenon occurs when the number of teeth of the pinion and the driven wheel has a common factor, i.e. an improper transmission ratio:

$$f_f = \frac{GMF}{H.C.M(z_1, z_2)} \tag{5.1.2}$$

In this situation, the teeth of the wheels are subjected to a cyclic contact of the same tooth. This configuration is harmful to the pair of gears that have a defect on the tooth side, as the meshing of the defective tooth will always weigh on the same teeth, reducing the useful life of the gear.

The last useful frequency for this analysis is the characteristic frequency of contact between the same pair of teeth:

$$f_r = \frac{GMF}{l.c.m(z_1, z_2)} \tag{5.1.3}$$

Generally, this phenomenon occurs at low frequencies and is difficult to detect.

To get a general overview of the frequencies involved in our case, the table below summarizes the values of the frequencies mentioned above, considering all gears of the gearbox evaluated for the different speeds taken into consideration during the test.

	Speed [RPM]	***Speed freq. [Hz]***	***GMF [Hz]***	***f_f [Hz]***	***f_r [Hz]***
Gear 1 (input)	900	15,000	315,000	105,000	0,938
	2100	35,000	735,000	245,000	2,188
	3000	50,000	1050,000	350,000	3,125
Gear 2	393,750	6,563	315,000	105,000	0,938
	918,750	15,313	735,000	245,000	2,188
	1312,500	21,875	1050,000	350,000	3,125
Gear 3	393,750	6,563	19,688	6,563	0,505
	918,750	15,313	45,938	15,313	1,178
	1312,500	21,875	65,625	21,875	1,683
Gear 4	30,288	0,505	19,688	6,563	0,505
	70,673	1,178	45,938	15,313	1,178
	100.962	1,683	65,625	21,875	1,683

Table 5.1-1:Frequencies of the gearbox.

Thanks to the obtained values, it is possible to identify the most significant frequencies spectrum components. This technique allows us to understand which frequencies elements are involved, identifying possible external frequencies and possible defects and/or reducer failures.

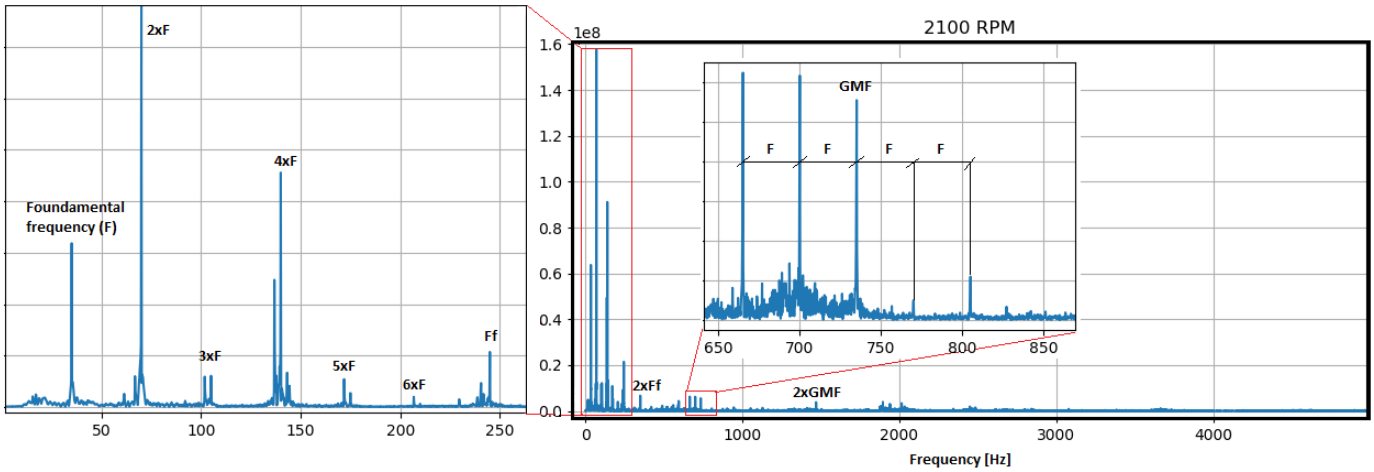


Figure 5.1.4: Individuation of the frequencies components relative to the reducer at the velocity of 2100 RPM.

As can be seen, in the case of a rotation speed of 2100 rpm it is easy to recognize the rotation frequency relating to 'Gear 1' (F) with all harmonics. The rotational frequency components of the other gears are not easy to spot. The latter has a lower weight than the 'Gear 1' so they take a back seat.
In *Figure 5.1.4* it is also possible to recognize the f_f relating to the first stage (Ff).
Finally, at higher frequencies, the GMF with the harmonics are found.
It is of particular interest to note that around the GMF some significant frequencies are reported, at a distance from the fundamental equal to the frequency relative to 'Gear 1'. These components are a clear example of signal modulation.
In the gear frequency analysis, it is common to encounter modulation phenomena mainly caused by wheel eccentricity and/or surface defects. These types of phenomena are evident in the spectrum and occur with lateral components with respect to the reference frequency, at a distance equal to the modulating signal.
However, generally, the side components of the modulated signal are symmetrical with respect to the main one. Instead, in the example above, the modulation components located to the left of the fundamental are much smaller than those located to the right. This phenomenon is due to an overlap of the effects and the irregularity of the waveform. In fact, by superimposing a variable load generated by the eccentricity of the wheel (which causes a modulation of the signal amplitude AM), with frequency modulation FM due to a periodic variation of the speed (caused by the clearance between the coupling of the toothing), we obtain a mixed modulation of the irregular waveform, which is identified with the asymmetrical conformation of the sidebands.
More precisely, the presence of excessive slack with respect to eccentricity manifest a decrease of the left sidebands.

these considerations can be explained through the frequency analysis of the signal modulation. Let us consider for simplicity a sinusoidal time-domain signal of amplitude A_T, frequency ω_T and phase Θ_T:

$$x_T(t) = A_T \cos(\omega_T t + \theta_T) \tag{5.1.4}$$

AM is a process based on the amplitude variation of the original signal (i.e. carrier signal) through a second signal (i.e. modulation signal). Considering a phase parameter equal to zero, the AM definition can be mathematically represented by the expression:

$$x_M(t) = \big(V_p + V_M \cos(\omega_m t)\big)\cos\,(\omega_p t) \tag{5.1.5}$$

Defining the modulation index, also called modulation depth, as the ratio of the amplitudes of the modulation signal and carrier signal

$$m = \frac{V_m}{V_p} \tag{5.1.6}$$

It is possible to rewrite the *expression 5.1.4* as:

$$\begin{aligned} x_M(t) &= (V_P + mV_P \cos(\omega_m t)) \cos(\omega_p t) \\ &= V_P \cos(\omega_p t) + mV_P \cos(\omega_m t) \cos(\omega_p t) \end{aligned} \tag{5.1.7}$$

Finally, considering the trigonometric formula of Werner, which allows obtaining the sum of trigonometric functions starting from the product of the same:

$$cos\alpha \; cos\beta = \frac{1}{2} \left[\cos(\alpha - \beta) + \cos(\alpha + \beta)\right] \tag{5.1.8}$$

We get the final formulation represented by the equation below:

$$x_M(t) = V_p cos(\omega_p t) + \frac{mV_p}{2} \cos(\omega_p - \omega_m)\, t + \frac{mV_p}{2} \cos(\omega_p + \omega_m)\, t \tag{5.1.9}$$

From this last equation, we deduce how the modulation results in the combination of three sinusoidal signals, shown in *Figure 5.1.5.b*, having as the main frequency ω_p and two components amplitude equal to $\frac{mV_p}{2}$ symmetrical to the carrier frequency with frequencies equal to the sum and the difference respectively, between the carrier and the modulating frequency [10].

Unlike amplitude modulation, frequency modulation is a kind of angle modulation that changes the phase and the frequency of the carrier signal. Moreover, in the FM it is not possible to obtain the FFT, as the FM produces an infinite quantity of symmetrical components with respect to the carrier frequency.
For simplicity, the case of FM for a periodic sinusoidal signal is considered. Taking *equation 5.1.4* as a carrier signal, the modulated frequency of the signal result:

$$x_{FM}(t) = \omega_p + \Delta\omega \cos(\omega_m t) \tag{5.1.10}$$

Where $\Delta\omega = A_p K_f$ represent the peak frequency deviation. Meanwhile, the phase of the modulated signal may be expressed as

$$\Theta(t) = \int_0^t \omega_i(\tau) d\tau = \omega_p t + \beta \sin(\omega_m t) \tag{5.1.11}$$

With $\beta = \Delta\omega/\omega_p$ that is the coefficient of the frequency modulation index of the signal.
For convenience, the resulting FM signal can be represented in the phasor notation[3]:

$$y(t) = Re\left\{A_p e^{j\omega_p t}\, e^{j\beta \sin(\omega_m t)}\right\} \tag{5.1.12}$$

[3] A phasor notation is a representation of a sinusoidal waveform in a single complex value that has the phase and the magnitude of the carrier signal $y(t) = Re\left\{Ae^{j\theta(t)}\right\}$.

Developing the Fourier series for the second exponential element we obtain

$$e^{j\beta \sin(\omega_m t)} = \sum_{n=-\infty}^{\infty} c_n e^{jn\omega_m t} \tag{5.1.13}$$

Substituting the mathematical expression for c_n and adopting a variable change for $\zeta = \omega_m t = \frac{2\pi t}{T}$, we obtain:

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j(\beta sin\zeta - n\zeta)} d\zeta = J_n(\beta) \tag{5.1.14}$$

In this way the *equation 5.1.13* become:

$$e^{j\beta \sin(\omega_m t)} = \sum_{n=-\infty}^{\infty} J_n(\beta) e^{jn\omega_m t} \tag{5.1.15}$$

Finally, we can obtain the expression of the AM signal represented by the equation below:

$$y(t) = Re\left\{A_p e^{j\omega_p t} \sum_{n=-\infty}^{\infty} J_n(\beta) e^{jn\omega_m t}\right\} = A_p \sum_{n=-\infty}^{\infty} J_n(\beta)\cos(\omega_p + n\omega_m)t \tag{5.1.16}$$

from the last expression, it is evident that the modulation of the sinusoidal signal has a theoretically infinite number of frequency components. These sidebands are also spaced from the carrier with a value equal to $\omega_p \pm n\omega_m$ for n-number of times function of the Bessel[4] degree and the value of the FM coefficient β[11,12].

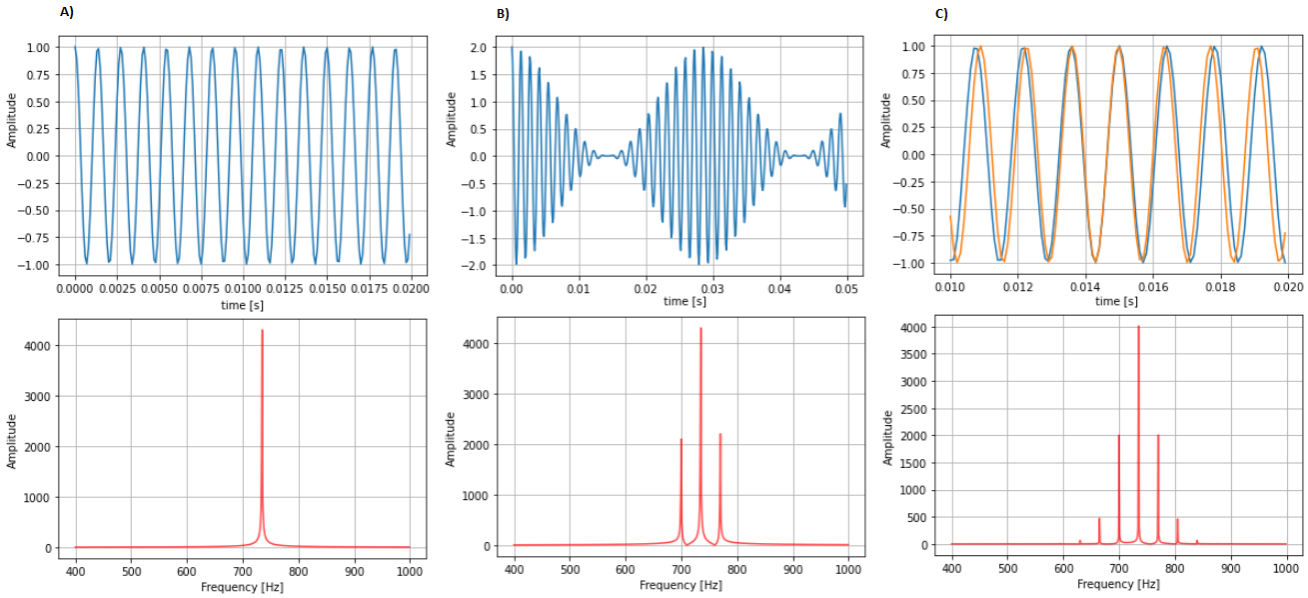


Figure 5.1.5: Representation of AM ad FM with the corresponding spectrum; A)carrier signal with frequency fp=735 Hz; B) AM applied to the carrier signal with Am=2, fm=35 Hz; C) Comparison between carrier signal (orange) and FM applied to the carrier signal (blue) with fm=35 Hz.

In the discussion reported above a sinusoidal signal has been considered.

[4] Bessel equations, which take their name from its creator Friedrich Wilhelm Bessel, are used In the field of mathematics to define cylindrical harmonics or Bessel functions.

In real applications, signal waveforms are not perfectly sinusoidal and usually, there are small variations in amplitude and frequency. Those variations, due to the previously mentioned construction irregularities, distort the sidebands, making them different from each other. Therefore, the superposition of the effects of AM and FM can characterize phenomena of destructive or constructive superposition of the signal, generating the spectrum reported in *Figure 5.1.4.*

These results were obtained by analysing different reducers. However, It is fair to say that the signal acquired by each gearbox is unique. Each reducer is characterized by distinctive sounds and frequencies studied up to now. Despite this, the construction and assembly phases introduce various factors that will characterize the sound and characteristics of the reducer.

5.2 Fault detection approach

With the previously discussed chapters, now it is possible to introduce various adopted methods for defective gearboxes detection. In particular, we have developed techniques able to bring out the periodic defect caused by surfaces irregularities present on the gears.
Since the "clicking sound" caused by one gearbox rather than another is different, several approaches based on completely different fundamentals have been developed. In this way, only the best solution found will be implemented in the final software.

The first method that we are going to analyse is based on the correlation theory of signals. This type of technique, also called serial correlation, consist of correlating a signal with a sample signal [13]. Using this method is possible to find where the two signals match.
The correlation between two discrete-time real signals (represented in our case by two audio waveforms) $x(k)$ and $y(k)$ can be written mathematically in the following form:

$$C_{xy}[k] = \sum_{m=-\infty}^{\infty} x[m]y[m-k] \tag{5.2.1}$$

In the case of autocorrelation, the two signals are the same, so $x[k] = y[k]$.

In our case, two approaches were considered: The autocorrelation of the signal and the correlation of the tested waveform with the defect filtered signal. In both cases, the results obtained are not optimal.
More precisely, for the first implementation, one would expect a result where a higher value appears periodically in correspondence with the 'tick'. Unfortunately, since the defect is not visible in the original waveform, the correlation does not find the values of the signal corresponding to the tick.
Instead, in the second approach, the waveform segment corresponding to the ticking was considered as a correlating signal (A series of frequency filters were used to isolate this sample). In this case, the correlation of the two signals was not very repeatable.

In fact, as previously mentioned, the signal relating to the tick is intrinsic to the reducer, manifesting itself with different waveforms and frequencies based on the rotation speed and the constructive characteristics of the reducer.
In conclusion, this solution was not too reliable due to the complexity of the type of signal. Furthermore, to proceed in a correct way it would have been necessary to think about a real-time extrapolation of the waveform related to the defect, weighing down the computational part of the program.

In the previous approach has been noted that the information provided by the signal in the time domain is not sufficient. For this reason, the second solution for the identification of the periodic defect has been conceived in such a way as to introduce the frequency component.

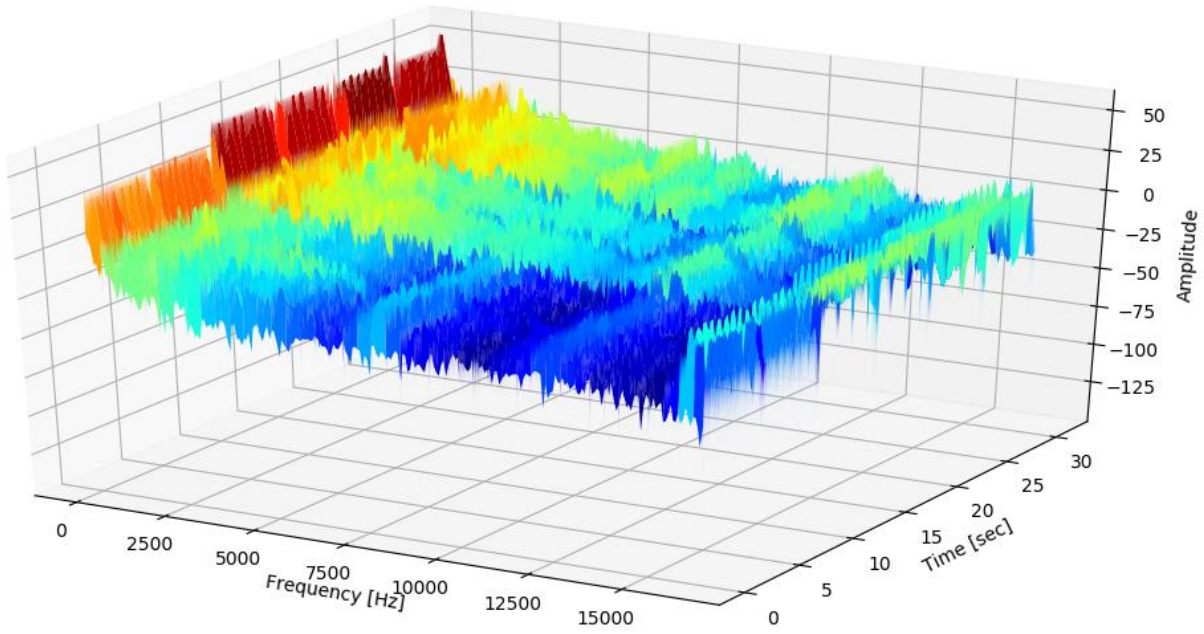


Figure 5.2.1: 3D plot representing the frequency components over time of a defective signal.

This representation could be seen as a result of the FFTs of the signal for each instant of time.
It is interesting to note that the frequencies over time are not very constant but follow a periodic behaviour. However, this comportment is not totally attributable to the ticking and can be attributed to signal modulation. In effect, this repetition is also present in the speeds where the defect is not present.
in any case, there are frequencies components between 2-5 kHz which represent the sound produced by the defect. Despite this, a threshold control on a large amount of data of this kind would be difficult and expensive, making even this type of approach impractical.

Therefore, we have seen how fundamental the introduction of both information on time and frequency is. For this reason, the following approach will maintain this characteristic but use a different visualization method, namely the spectrogram.
The frequency components seen previously, corresponding to the defects, are translated into visible vertical lines in the spectrogram, as shown in *Figure 5.2.2.*

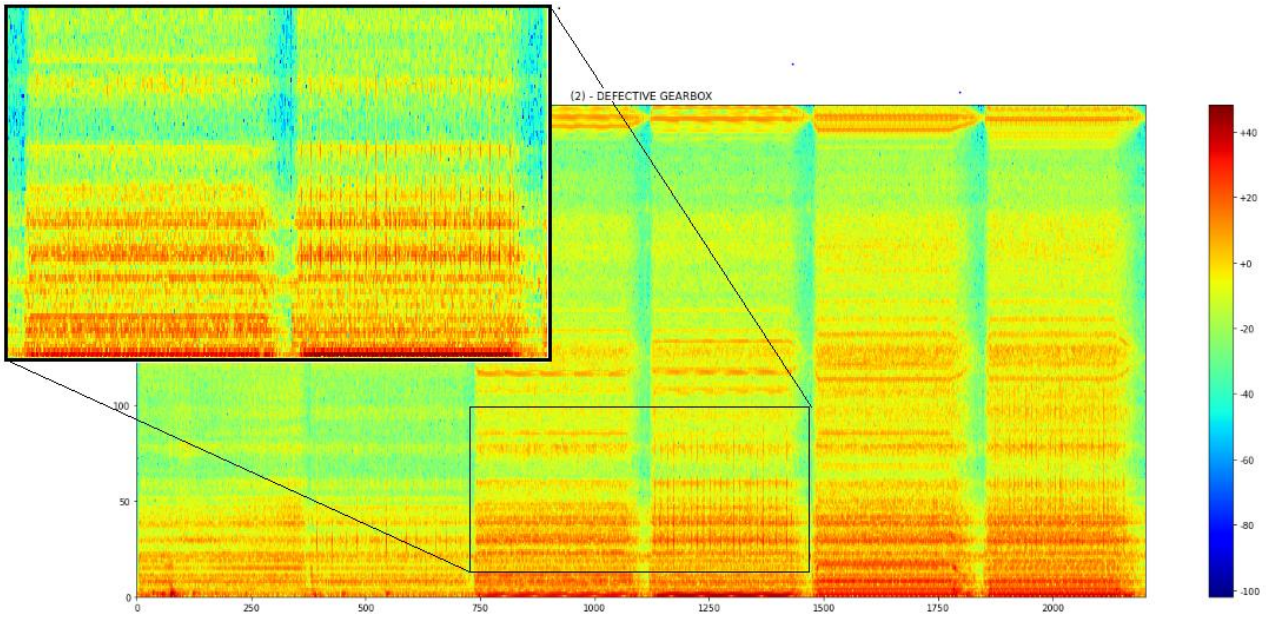


Figure 5.2.2: Spectrogram of a defective reducer in counterclockwise speeds.

above an evident example of a defect has been reported, but the spectrogram is not always so clear. Furthermore, the zoom shown above allows to highlight the same speed evaluated in both directions. It is evident how the speed reported in the right (counterclockwise direction) presents the discussed vertical elements attributable to the ticking.

I order to identify the presence of these vertical elements in the spectrum a technique was devised and reported in *Figure 5.2.3*. More precisely, the test is divided by speed, removing the acceleration and deceleration chunk to compute the spectrogram. Then, each bin of the spectrogram was subjected to an FFT to find the repetition frequency of the elements. Subsequently, the result obtained is filtered with a comb filter leaving unchanged the frequency on interest (frequencies related to gear 2) and its harmonics, cancelling the other components.
In the end, the result obtained is used to compute the percentage component of the power spectrum covered by the ticking to respect the total one. The result is subsequently passed through a threshold control which catalogues defective or good the bin under control. If the percentage is high, then the analysed bin presents elements with a repetition frequency equal to the frequency rotation of Gear 2. Therefore, a defect is present.
Of course, it can happen that a pattern repeated at the same frequency as the defect is detected (for example modulation at low frequencies). To avoid this situation, another threshold control has been implemented in the algorithm.

This last step can signal the next, or almost successive bins (some bins of the defective signal may not present repetitive patterns, this situation is due to the lack of visibility of some vertical elements), in which the defect is detected.

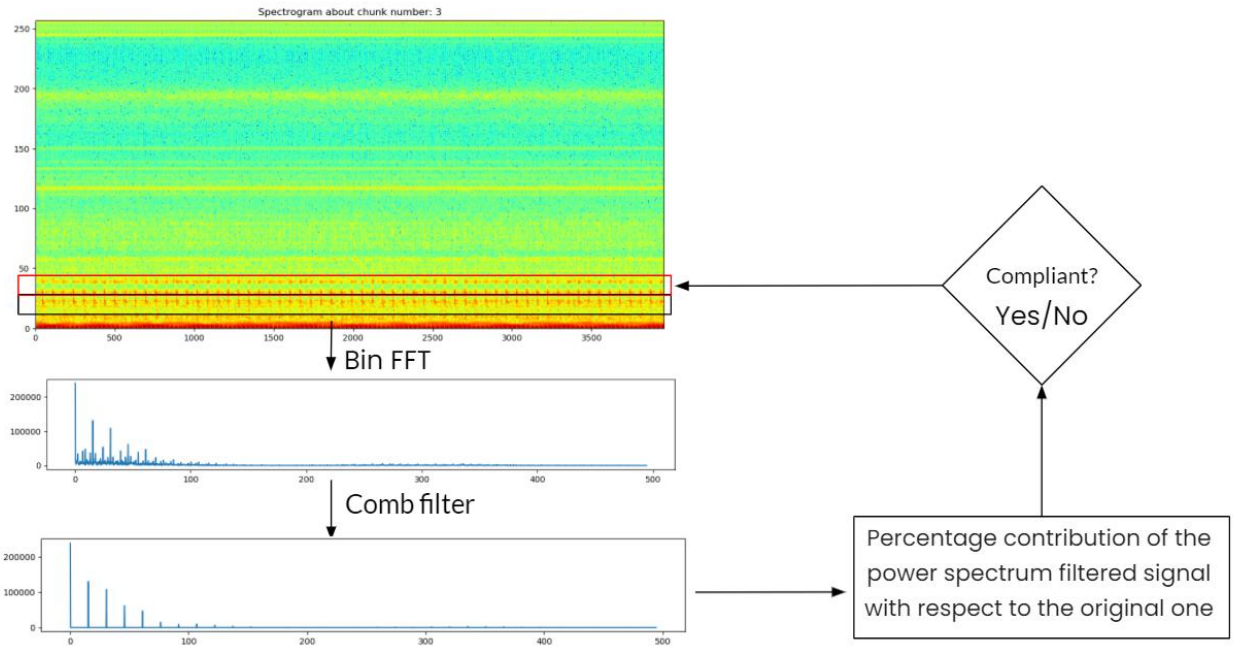


Figure 5.2.3:Defect detection procedure.

With this method, it was possible to achieve high precision in identifying faulty gearboxes. Even in those reducers where the noise generated was very slight.
Given the excellent effectiveness, this approach has been implemented directly in the recording and data collection software. Thanks to this addition, the data can be catalogued and saved automatically in the corresponding category.

5.3 Noise detection approach

As mentioned at the beginning of the paper, the second problem encountered is relative to noise. Although this problem has already been dealt with in the past, today, it affects a small percentage of gearboxes.
In the past batches, it was found that the wear of the gear hobbing caused an increased noise level of the reducer. Despite this, it is not explained why there are few noisier gearboxes within the last lot.
The identification of the latter will be indispensable for the restoration and the analysis by the mother company, which is fundamental for improving the product quality.

For these reasons, the frequency spectrum of the signal is controlled to identify possible anomalies. This system control is based on the statistical study of gearboxes classified as 'good'. More precisely, thanks to the data collection, it was possible to create an average spectrum related to the non-noisy gearboxes. Subsequently, to take into account the diversity of the sounds produced by the reducers, and to avoid too stringent conditions, a value equal to three times σ (*equation 5.3.1*) was added to the average

value of the spectrum. In this way, we obtain a threshold over the entire frequency spectrum.

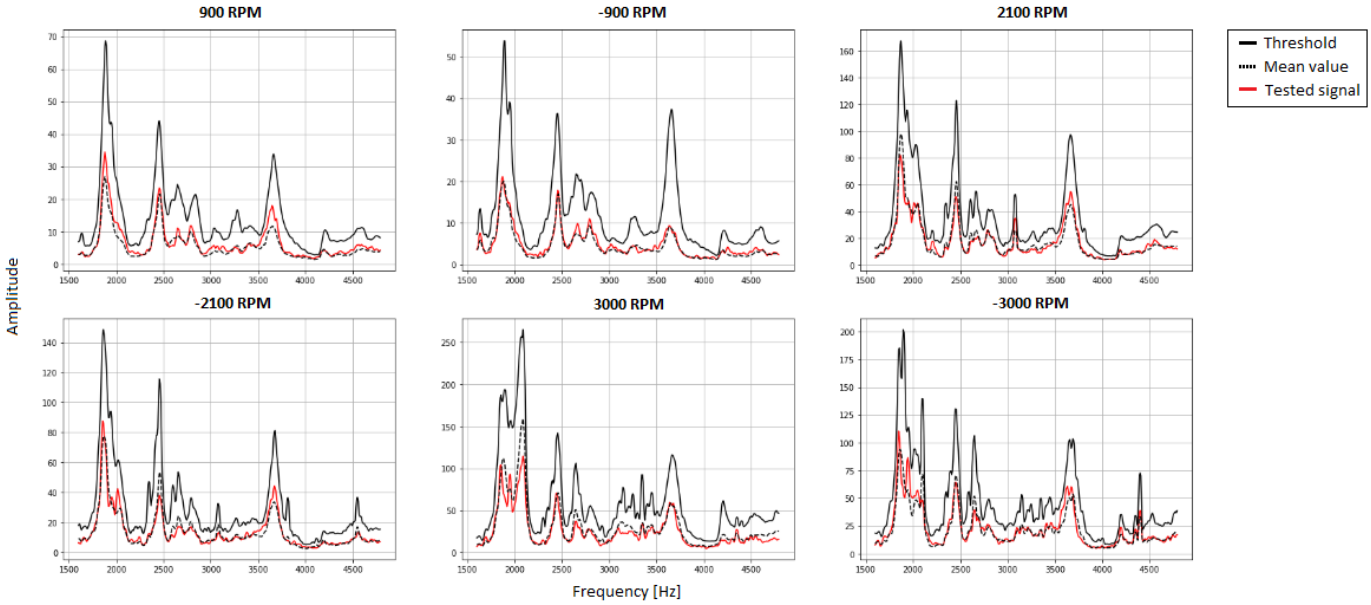


Figure 5.3.1: Noise threshold for the speed tested.

From the figure above, the threshold value, obtained as the sum of the black dotted line plus 3 times the variance of the dataset, is placed above the value of the tested gearbox (red line). In this case, therefore, the gearbox is acceptable and can continue to the assembly phase.
For display purposes, only the critical frequency range has been reported.

The choice to take as an additional value equal to 3 times the variance of the population is the result of a trial-and-error process. Thanks to the feedback from the production department, the gearboxes that did not fall within the thresholds were evaluated after assembly. The value obtained from this process is equal to:

$$\sigma = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x}) \qquad (5.3.1)$$

Where x_n represent the individual values in the dataset, $\bar{x}$ is the mean value of the dataset and N is the number reducer present in the dataset.
This last parameter was assumed to be equal to the number of reducers contained in a batch. This consideration allows to have a history of the noise levels of each batch. in this way, direct comparison between the previous batches will be possible. In addition, this strategy will allows to keep quality under control, improving it as much as possible.

6 Achieved results

The results obtained in this first part of the thesis are comforting.
The developed system (reported in *Figure 6.1*) is able to detect, catalogue and notify non-compliant gearboxes with high precision.

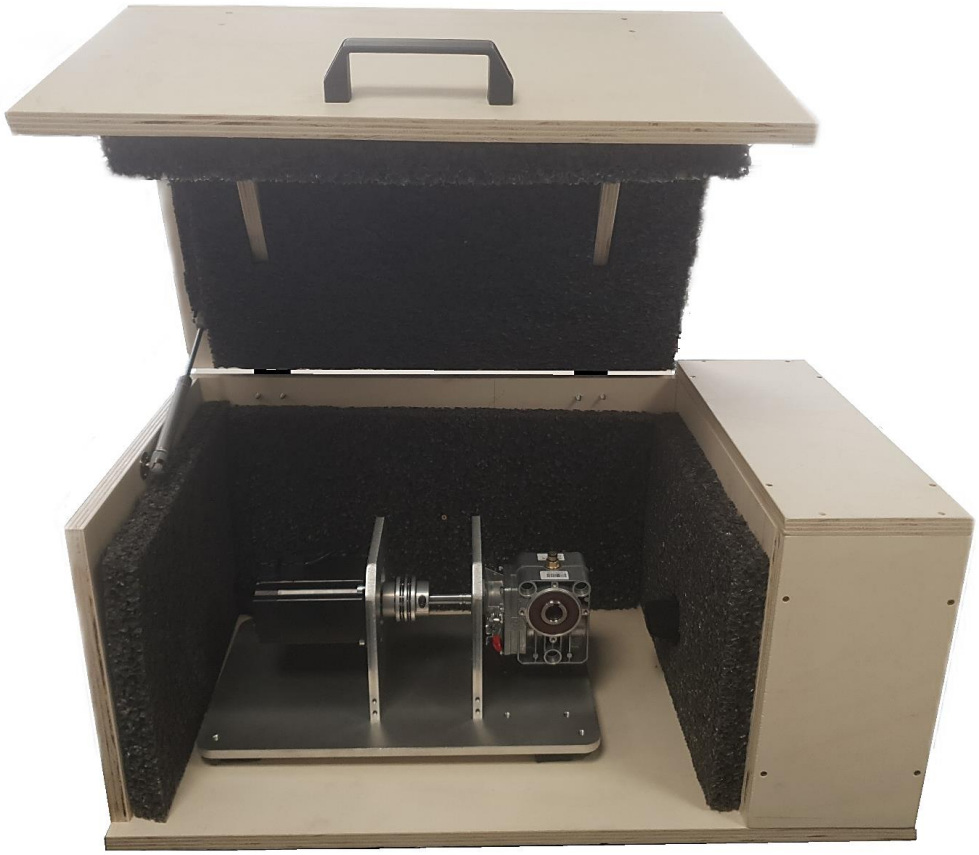


Figure 6.1: Test bench

The implementation of this technology allows the company to have an automatic system for the verification and control of the purchased gearboxes. This configuration is able to avoid the large waste of material in the assembly process of the DMD0 motors.

The methodologies used and refined during this research made it possible to identify even the most difficult gearboxes to judge.
Furthermore, thanks to the results obtained, it will be possible to supply the equipment to the manufacturer of the gearboxes. In this manner, a product quality improvement process may be followed.

Despite these excellent results, it was decided to develop another classification software that differs from the classical programming approach.
With a view to a future onboard implementation for predictive analysis and maintenance, we wanted to develop artificial intelligence software to detect and classify anomalous events.

7 Introduction to Artificial Intelligence

In 1956 the term artificial intelligence (AI) was coined during the summer conference at Dartmouth college by John McCarthy[14]. He defined AI as *"the science and engineering of making intelligent machines".*

Nowadays, with continuous innovations and research of the last year, the AI definition has been revisited in a modern key as *"the study and design of intelligent agents"*, where intelligence agents are intended for systems (Human, Robotic, Software etc.), able to study the environment and take the best decision to maximizes the success probability.

J. Stuart and P. Norvig in the publication of their book (*Artificial Intelligence: A Modern Approach*.), report four historical approaches to define AI, namely[15]:

1. Thinking humanly
2. Thinking rationally
3. Acting humanly
4. Acting rationally

It is possible to quickly illustrate these four categories by defining the ideas of thought as processes of reasoning and identification of a solution, while the ideas of action are strictly correlated with the behaviour of the solution found. Furthermore, these ideas may reflect human or, as in the case of interest, rational behaviour/reasoning.

All these approaches can define the various fields in which AI operates, such as Intelligent data processing, virtual assistant, physical solution, voice generation and much others.

we have seen how the term artificial intelligence covers a huge field of applications and solutions. Over the years, the concept of AI, especially that referred to "acting rationally", has increasingly developed in the sub-categories, reported in *Figure 7.1*.

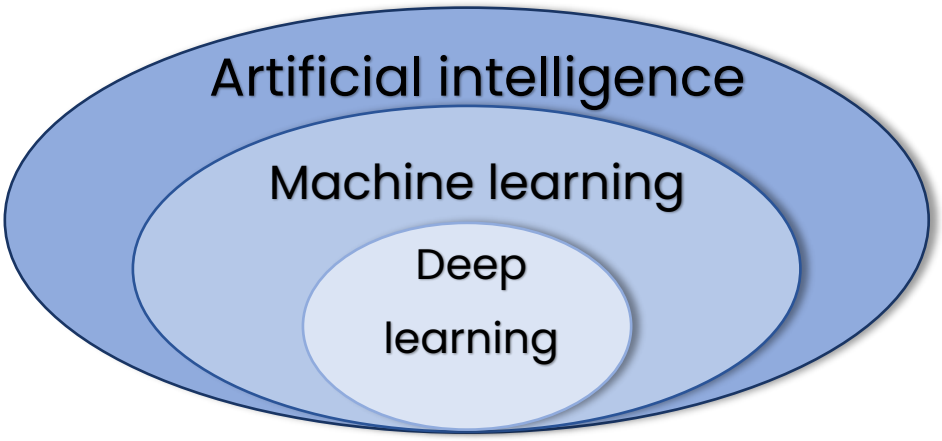


Figure 7.1:AI subfields

Machine learning (ML) is a subfield of AI. This philosophy is based on algorithms able to learn some features without the aid of specific programmed commands. Deep learning (DP), in turn, is a subset of ML. the substantial difference concerning the ML is that the DL uses a vast neural network capable of extrapolating from the data the features necessary for training, avoiding features engineerization.

7.1 Introduction to Deep Learning

In this second part of the thesis, we focus on DL applied to the sound in order to classify, in an automatic way, the defective gearboxes and the good ones. We have previously introduced DL as a subset of AI.

In this chapter, we will introduce the basic knowledge that governs this technology. To have a general overview, it is important to discuss the first neural network created, called Perceptron.

7.1.1 Perceptron

The model of Perceptron was designed by Rosenblatt in 1958 with the scope to have a binary classifier able to do predictions based on a linear prediction algorithm that combines weights and features [16]. The model, reported in *Figure 7.1.1.1*, is very simple and is based on a few elements: a series of inputs (x_i) to acquire the input signal, a sinaptic weights (w_i) to ponderate the input signal, a summing junction that sums all the weighted inputs

$$v_k(n) = \sum_{i=0}^{m} w_i(n)x_i(n) \tag{7.1.1.1}$$

and an activation function (a) that applied a step rule to retrieve the output:

$$y_i(n) = a(h_k(n)) \tag{7.1.1.2}$$

Normally as activation function is used a sigmoid function is represented by the following equation:

$$a = \sigma(h) = \frac{1}{1 + e^{-h}} \tag{7.1.1.3}$$

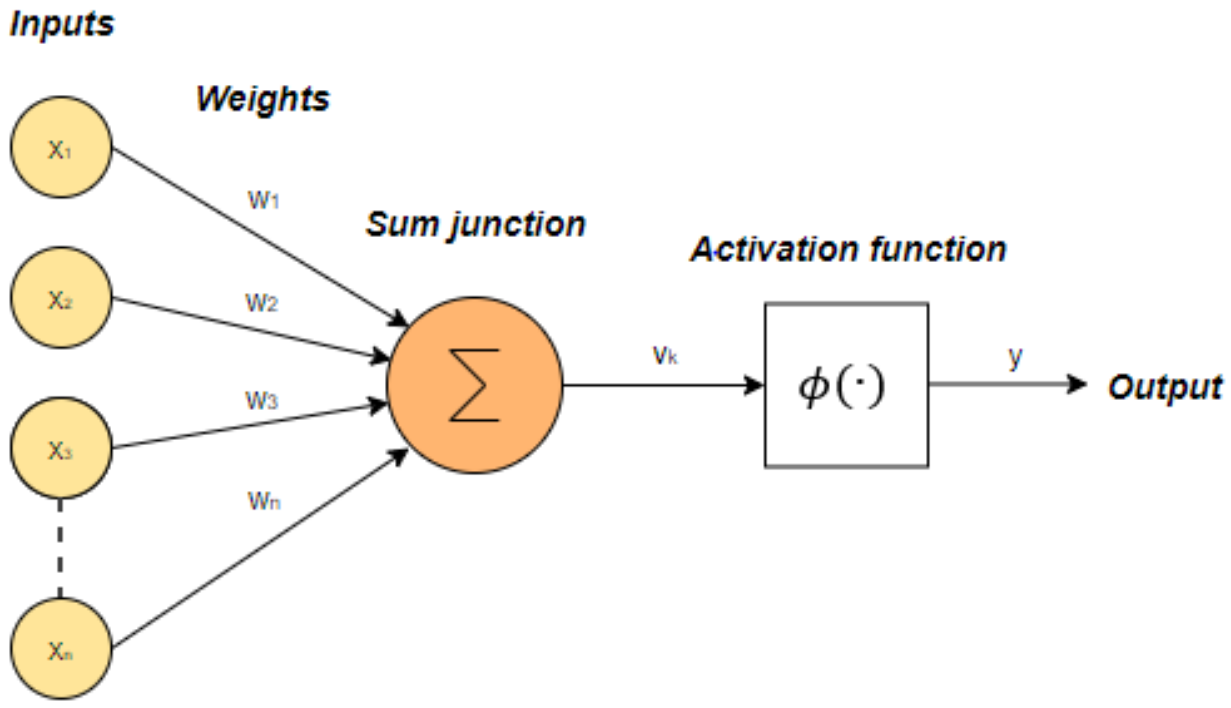


Figure 7.1.1.0.1: Perceptron model.

The reported figure represents a single perceptron model that is the key point for every neural network architecture. From this structure, it is possible to distinguish two main types of architecture:

- Single/multi-layer feedforward networks, in which the flow of information proceeds only in the direction of the output;
- Recurrent networks, where feedback is present to influence the input from the previous output results.

From those main types of neural networks is possible to derive other typologies with specific characteristics depending on the application [17].

7.1.2 Neural network learning

Like the human brain, a neural network is capable of learning and interacting with the environment. In this context, the term learning is defined as the action/reaction of a certain situation, with the aim to be able to replicate a particular behaviour in the presence of a given input.

In the case of multilayer feedforward networks, reported in *Figure 7.1.2.1*, the learning phase consists to take and process an input vector from the training samples, getting a prediction and computing the error. Then, If the output obtained by the network does not correspond with the aspected value, the connection w_i will be modified to achieve the correct value.

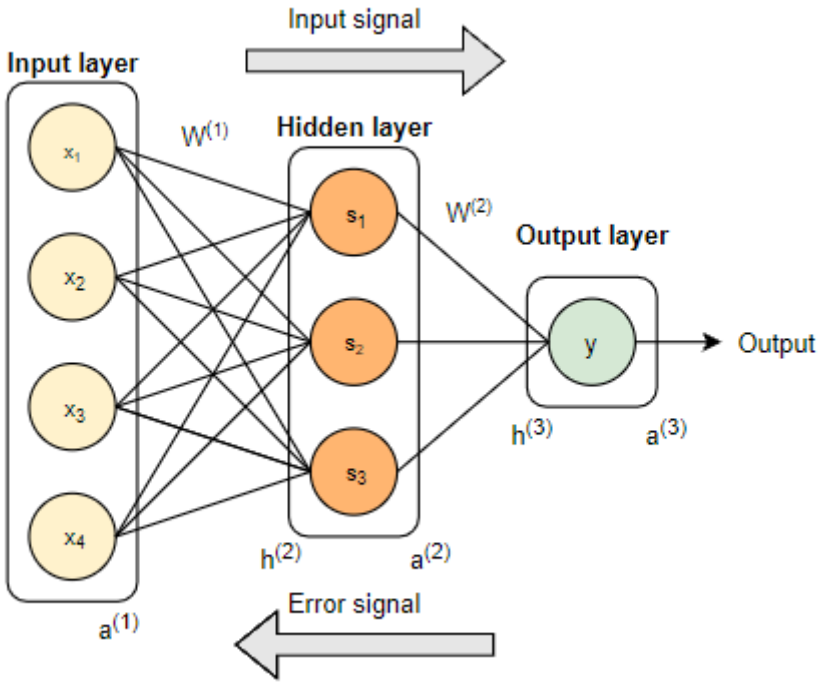


Figure 7.1.2.0.1: Feed-forward Neural Network with a single hidden layer and 3 neurons.

This process is strictly correlated with the minimization problem of the difference between the expected value wrt the real one (loss function ζ). This optimization problem is based on the backpropagation of the error. in particular, the error return to the input layer in order to change and adjust the value of the weights and compute the new output. In other words, the backpropagation algorithm results in a correction of the

synaptic weight $\Delta w_i(n)$, proportional to the partial derivative of the loss function $\left(\frac{\delta\zeta}{\delta w}\right)$, i.e. the gradient of the error function. This information represents the sensitivity factor, which determines the right direction of search of the local minimizer, improving the predictions [18].

This process is repeated n-times, i.e. n-epochs, until the value obtained converges with the expected one. In this case, the partial derivative is approximated equal to zero.

After this brief introduction, we can mathematically analyze this concept. Let's consider an input signal that propagates to the subsequent layer until the final level, i.e. the prediction layer. Here the error is computed as a function of the prediction y, and the expected output y_e.

$$\zeta = \zeta(y, y_e) = \frac{1}{2}(y - y_e)^2 \tag{7.1.2.1}$$

In this specific case, for simplicity, we have used a quadratic error function.

The next step, as mentioned above, is the computation of the gradient of the loss function so de partial derivative of ζ regard all the weights W^n[5]. To achieve this result, we consider the neural network as a very complex function that depends on the input and the weight values:

$$F = F(x, W) \tag{7.1.2.2}$$

Based on this information we can compute the loss function playing the chain rule:

$$\frac{\delta\zeta}{\delta W^{(2)}} = \frac{\delta\zeta}{\delta a^{(3)}}\frac{\delta a^{(3)}}{\delta h^{(3)}}\frac{\delta h^{(3)}}{\delta W^{(2)}} \tag{7.1.2.3}$$

Remember the *equation 7.1.2.1* where $y = a^{(3)}$ the first term result as

$$\frac{\delta\zeta}{\delta a^{(3)}} = 2\frac{1}{2}\left(a^{(3)} - y_e\right) = a^{(3)} - y_e \tag{7.1.2.4}$$

Then, considering as an activation function (a) the *equation 7.1.1.3*, the second term of *equation 7.1.2.3* is equal to:

$$\frac{\delta a^{(3)}}{\delta h^{(3)}} = \sigma'\left(h^{(3)}\right) = \sigma\left(h^{(3)}\right)\left(1 - \sigma\left(h^{(3)}\right)\right) \tag{7.1.2.5}$$

Finally, the last term is represented by:

$$\frac{\delta h^{(3)}}{\delta W^{(2)}} = \frac{\delta\left(a^{(2)}W^{(2)}\right)}{\delta W^{(2)}} = a^{(2)} \tag{7.1.2.6}$$

Now we have all components for rewriting the *equation 7.1.2.3* as

$$\frac{\delta\zeta}{\delta W^{(2)}} = \left(a^{(3)} - y_e\right)\sigma'\left(h^{(3)}\right)a^{(2)} \tag{7.1.2.7}$$

[5] W^n are the weight matrices that represents the connection between the different neurons with the different layers $w_{n,l}$.

In this way, we have obtained the equation of the loss function relative to the second layer. The same demonstration can be adopted for the gradient of the first layer. The result, as reported in *equation 7.1.2.7*, is the combination of the loss function gradient of the second layer with the addition of other terms relatives to the first layer:

$$\frac{\delta\zeta}{\delta W^{(1)}} = \frac{\delta\zeta}{\delta a^{(3)}}\frac{\delta a^{(3)}}{\delta h^{(3)}}\frac{\delta h^{(3)}}{\delta a^{(2)}}\frac{\delta a^{(2)}}{\delta h^{(2)}}\frac{\delta h^{(2)}}{\delta W^{(1)}}$$
$$= \left(a^{(3)} - y_e\right)\sigma'\left(h^{(3)}\right)W^{(2)}\sigma'\left(h^{(2)}\right)x \qquad (7.1.2.8)$$

Therefore, with this mathematical proof, we can notice how the error calculated at the end of the neural network propagates in the previous layers.
Now, knowing the gradient relative to all the layers, we can apply a minimization function to find the best weight for all layers. In particular, the gradient descent algorithm is usually used to minimize the loss function. To reach local minima the minimization function takes a step in opposite direction to the gradient. The step used to move towards the optimal minimizer is governed by a parameter called the learning rate.

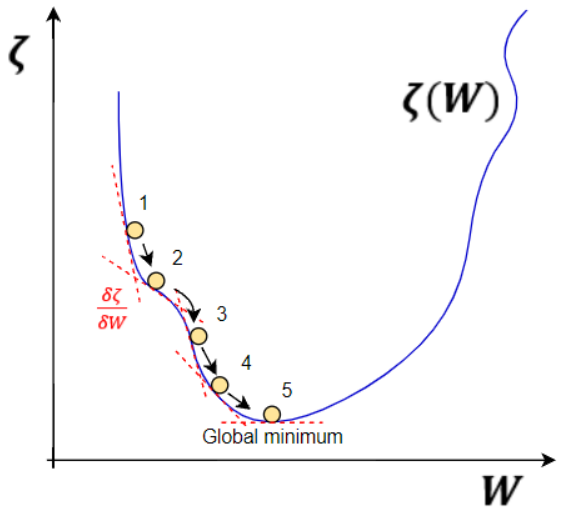


Figure 7.1.2.0.2: Gradient descent algorithm.

The choice of the learning rate is fundamental. A small value would lead to a very slow to found a solution. Conversely, a too large value would lead to the divergence of the solution.

7.2 Convolutional neural network

The application treated in this thesis provides for the classification of audio files. This category of files is characterized by a large amount of data, just think that using a sampling frequency of 44100 Hz, one second of recording is equivalent to 44100 input data. This large amount of information is difficult to manage and process with moderate training and execution times.
Furthermore, the data obtained from the raw signal are not very characterizing in the identification of defects. For this reason, it was decided to work on an image dataset. More precisely, on the spectrograms obtained from the recordings.

In order to work with images as input, a specific neural network must be used. This neural network should be able to extrapolate the features present in the pictures, replicating the perception of the human eyes.
For this scope, a convolutional neural network (CNN) is introduced.

CNN or ConvNet is a feed-forward multilayer neural network characterized by the implementation of convolution layers, which are extremely important to extrapolate the main features of the image. Unlike a classic neural network, CNN uses a series of filters able to focus on the details of the image and not on the overall. This feature, in our case, makes the use of this neural structure perfect, as the goal is to identify those vertical lines characteristic of the 'ticking'. despite this, NCCs usually use multiple layers such as:

- Convolutional layers: based on the convolution of vectors defined by the following equation:

$$\boldsymbol{y} = \boldsymbol{x} * \boldsymbol{w} \rightarrow y_i = \sum_{k=0}^{m-1} x_{i+m-k} w_k \tag{7.2.1}$$

  Where $\boldsymbol{x}$ represent the input vector, $\boldsymbol{w}$ is the kernel with size n and m respectively. In our case, the input is represented by an image, so the input results in a matrix of pixels, then the principle is redefined in a 2D convolution, and the equation 7.2.1 becomes:

$$\boldsymbol{Y} = \boldsymbol{X} * \boldsymbol{W} \rightarrow Y_{ij} = \sum_{k_1=0}^{m_1-1} \sum_{k_2=0}^{m_2-1} X_{i+m_1-k_1, j+m_2-k_2} w_{k_1,k_2} \tag{7.2.2}$$

  With matrix $\boldsymbol{X}$ and $\boldsymbol{W}$ dimension n_1 x n_2 and $m_1$and m_2 respectively.
  In other words, the kernel (digital filters) are scrolled on the image to obtain a weighted result, which will be used as input for the next layer. *Figure 7.2.1* shows the convolution layer principle using a 3x3 kernel on a 5x5 image. The result obtained is placed in the next layer in the centre of the area taken into consideration, then the focus area slides to the next position to calculate the new output. The output dimension matrix is strictly correlated with the size of the kernel and the stride, i.e. the slide step of the area of interest, in the horizontal and vertical directions.

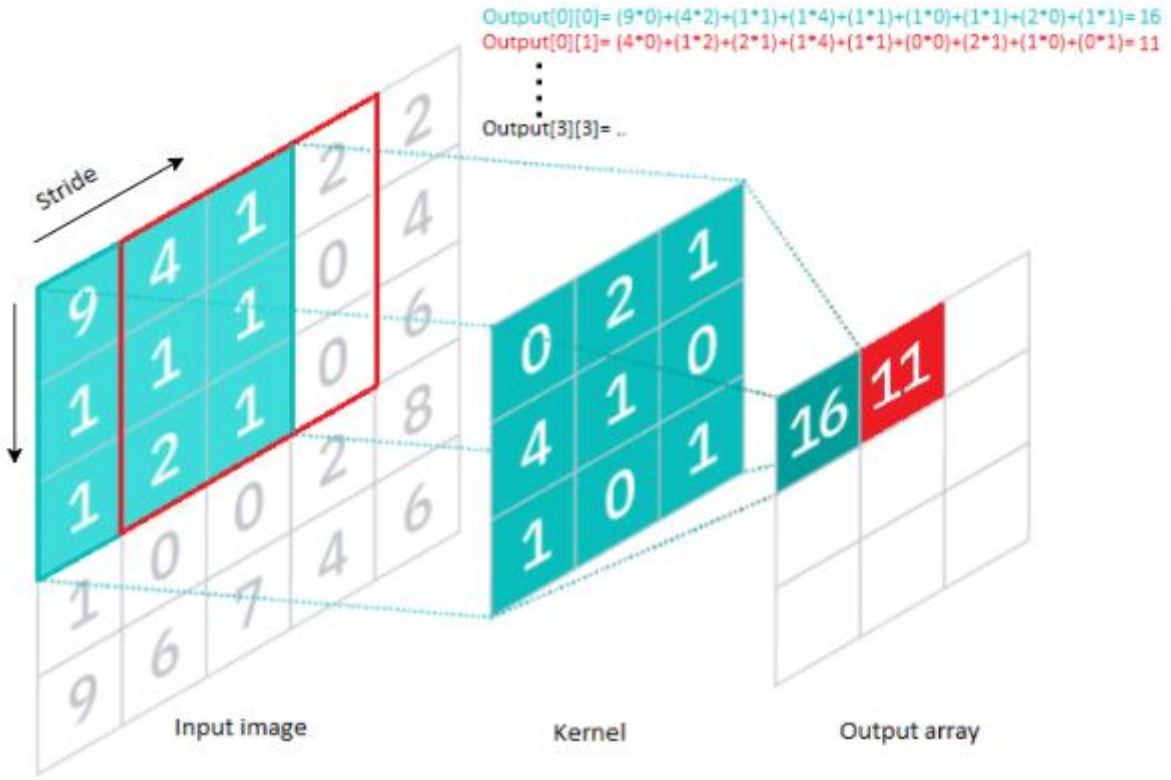


Figure 7.2.1: Convolution layer principle.

- Nonlinearity: After each convolutional layer a Rectify Linear Unit (ReLU) is applied to avoid non-linearities in the model. ReLU transformation can be modelled as:

$$ReLU = \begin{cases} 0, if\ x < 0 \\ x, if\ x \geq 0 \end{cases} \quad (7.2.3)$$

- Pooling layer: called also downsampling, introduces a resize of the input. An unweighted filter of dimension n x n, typically 2x2, is applied to obtain a smaller output dimension. Two main filters were moved across the input:
  - Max pooling: is the most used and return the pixel with the maximum value.
  - Average pooling: returns the average of the pixel contained in the kernel.

  This process introduces an elevated loss of information, but at the same time, decrease drastically the complexity of the problem, improving the neural network efficiency.

- Fully-connected layer: is used to classify the input. This layer connects all of his neurons with the neurons of the successive layer. Usually, to retrieve the probability of output correctness a Softmax function is applied. The softmax equation is given by:

$$o_i = \frac{e^{z_i}}{\sum_{i=1}^{M} e^{z_i}} \quad (7.2.4)$$

  Where O_i is a number between 0 and 1 that represent the probability of the ith number, z_i is the output i before the application of the softmax equation and M is the number of total outputs[19].

7.2.1 State of the art

Since 1998, with the introduction of LeNet, the convolutional neural networks become very interesting in the field of recognition and classification of images, starting the era of CNN. This type of structure, with the implementation of a subsampling layer, introduced a significant reduction of parameters at the same depth of a multi-layer network. LeNet was introduced for the classification of handwriting numbers. This net, was the first network that combines the sequence of convolution, pooling and non-linearity layers, becoming the key features of CNN. However, this structure was characterized by one main problem, i.e., the possibility to have a very low gradient in the backpropagation, causing, in some cases, the interruption of network training.
With the introduction of new datasets and competitions, many other CNN architectures have been developed, such as:

- AlexNet: Introduced in 2012, consist of similar architecture as LeNet with a deeper hidden layer. this structure uses a series of filters of different sizes (11x11, 5x5, 3x3), a data augmentation to avoid overfitting, activation of ReLu after each layer of convolution and max pooling.

- ZFNet: winner of ILSVRC 2013 competition. It was achieved by adding elements of deep learning and modifying the hyper-parameters of the AlexNet network.

- GoogLeNet: also called Inception, achieved an excellent result in 2014, reaching a very close performance of the human eye. Its architecture introduces 22 layers based on very small convolution modules implemented to reduce the number of trainable parameters. Also, this structure implements the inception module, which is a structure capable to process the input vector with three filters of different sizes (1x1, 3x3, 5x5). Then, The result obtained is concatenated and sent to the next layer.

- VGGNet: uses a uniform and simple architecture, which is why it is used in many features extraction applications. In addition, the weight configuration has been made public and available to the community.

- ResNet: Residual Neural network, introduce a skip connection to pass directly to successive layer. This solution avoids the problem of vanishing gradient and drastically simplify the structure, skipping some layers in the initial training phase.

Before this introduction and a brief history of CNN, only one structure will be deep discussed in this thesis.

7.2.2 VGG16 Architecture

In this application, we have chosen to use a network capable of achieving a good percentage of accuracy with a time of execution and training not too high. For this purpose, CNN VGG16 has been implemented.
The architecture reported in *Figure 7.2.2.1* was proposed by K. Simonyan and A. Zisserman for the *"ImageNet Large Scale Visual Recognition Challenge 2014"* (ILSVRC2014)[20].

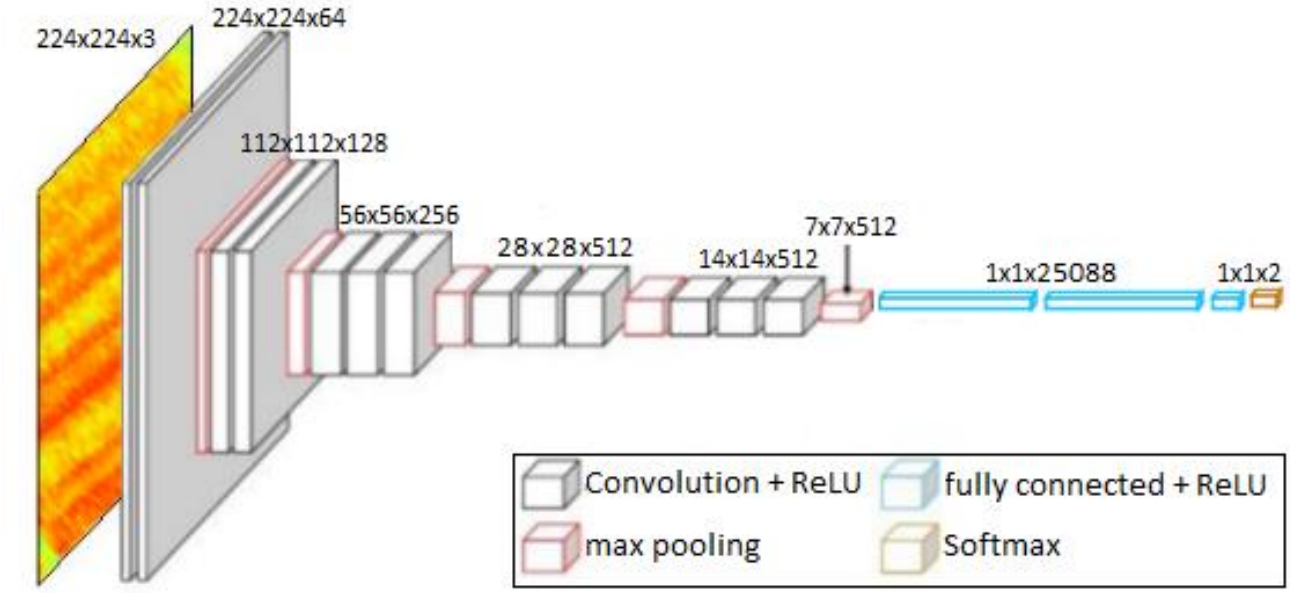


Figure 7.2.2.1: VGG16 neural network architecture.

The very deep architecture of this network allows you to extract the various features through 16 layers (13 convolutions and 3 fully connected). The input image is resized so that it has an RGB format of 224x224 size.

The very deep architecture allows the neural network to extract the various features through 16 layers (13 convolutions and 3 fully connected). The input image is resized so that it has an RGB format of 224x224 size, which is processed through the first group of convolution layers with a kernel of 3x3 size and stride equal to 1 pixel. Subsequently, the format of the matrix is reduced in half performing a Max-pooling over 2x2 pixel mask with stride equal to 2.

the latter operation is performed later for the remaining convolution blocks up to the fully connected layers with different depths, 4096 for the first two and 2 for the last, which corresponds to the number of outgoing classes.
Important to remark is that a non-linear rectifier (ReLU) was used for each hidden layer.

7.3 Dataset and training

For the realization of the final model, two different approaches have been adopted: The first one is based on transfer learning, therefore, on a fine-tuning of the pre-trained model. Instead, the second approach is based on the expansion of the dataset for training the network from scratch.

Due to the few cases of defective gearboxes compared to those judged good, the dataset turns out to be too small. For this reason, it was thought to use a complete dataset based on 4,197,122 annotated images according to the WordNet hierarchy, and subsequently refine the network for our specific case, providing spectrograms obtained from the recordings as inputs. In this way, the number of samples increases and the image resolution remains better. Subsequently, to obtain the right format required by the neural network, the spectrogram is resized, cutting the frequencies in which the defect was not present, i.e., the high frequencies, obtaining a 224x224 RGB image. The result of this process is represented in *Figure 7.3.1.a*.

In the second case, to increase the number of defective samples, a new registration for defective cases has been carried out. In addition, thirty seconds of record are saved to be split into new twenty samples. Unfortunately, this method was not sufficient.

For this reason, a fake dataset was generated in order to replicate the periodic defective. For this issue, the spectrogram of a wide-band periodic signal, (with the same period of the defect), was superimposed on the spectrogram of a 'good' gearbox.

Subsequently, the generation process was automated. The introduction of random parameters allows the replication of a wide range of situations. In particular, it was possible to modify the frequency band of the signal, the initial position of the periodic sound, the amplitude and the intensity of the noise. An example of the obtained results is reported in *Figure 7.3.1.b*.

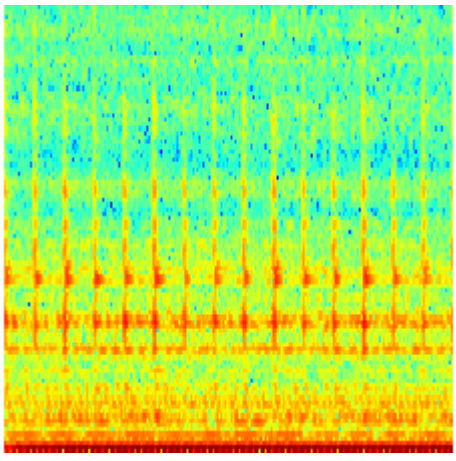


Figure 7.3.1.a: Spectrogram of the defective gearbox

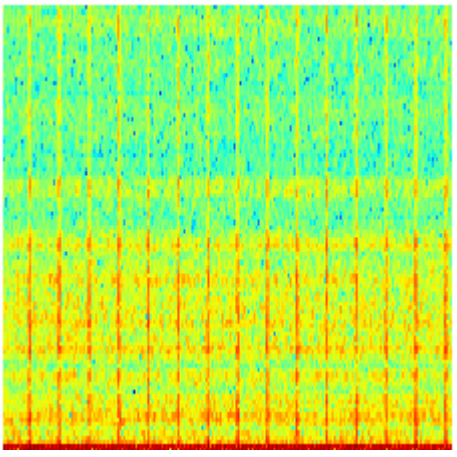


Figure 7.3.1.b: Auto-generated fake spectrogram.

Furthermore, during the training, the dataset was extended with an augmentation that consists of an image generator function, implemented in python, able to process the width, height and shear shift range.

7.4 Software implementation

Also in this case, the software was implemented using the Python programming language. In particular, Keras, an open-source library for ML, was imported. This library, released by F. Chollet in 2015, is a user-friendly tool that acts as an interface with Tensorflow.

Keras implement directly all the necessary layers, models and optimizers, fundamental for the realization of the model. In our case, the standard VGG16 architecture is just implemented and can be imported as *tf.keras.application.vgg16*. In order to compile the model, different parameters are specified. Among these parameters, we can mention: the loss function, defined as binary cross-entropy, governed by the following equation:

$$\zeta = -\frac{1}{len(y)} \sum_{i=1}^{len(y)} y_{e(i)}\, log\, y + \left(1 - y_{e(i)}\right) \log(1 - y) \tag{7.4.1}$$

Usually, this objective function is applied to classification problems as in the considered application; The input data, that for this structure the best choice is a matrix of size 224x224x3, where the last number refers to the RGB mode; The output size, equal to 2 that is the number of class of the classifier (['Godd'; 'Defective']) and, in the end, the number of epochs, i.e., the training and test iteration among the entire dataset.

Another important parameter is represented by the initialization of the weights of the neural network. As discussed in the previous chapter, two approaches are considered. For the implementation of transfer learning the 'weight' parameter is set equal to *imagenet,* which refer to the homonymous dataset. In the other case, weights initialization is random.

Layer (type)	Output shape	Param #
input_1 (InputLayer)	(224,224,3)	0
block1_conv1 (Conv2D)	(224,224,64)	1792
block1_conv2 (Conv2D)	(224,224,64)	36928
block1_pool (MaxPooling2D)	(112,112,64)	0
block2_conv1 (Conv2D)	(112,112,128)	73856
block2_conv2 (Conv2D)	(112,112,128)	147584
block2_pool (MaxPooling2D)	(56,56,128)	0
block3_conv1 (Conv2D)	(56,56,256)	295168
block3_conv2 (Conv2D)	(56,56,256)	590080
block3_conv3 (Conv2D)	(56,56,256)	590080
block3_pool (MaxPooling2D)	(28,28,256)	0
block4_conv1 (Conv2D)	(28,28,512)	1180160
block4_conv2 (Conv2D)	(28,28,512)	2359808
block4_conv3 (Conv2D)	(28,28,512)	2359808
block4_pool (MaxPooling2D)	(14,14,512)	0
block5_conv1 (Conv2D)	(14,14,512)	2359808
block5_conv2 (Conv2D)	(14,14,512)	2359808
block5_conv3 (Conv2D)	(14,14,512)	2359808
block5_pool (MaxPooling2D)	(7,7,512)	0

Layer (type)	Output shape	Param #
input_1 (InputLayer)	(224,224,3)	0
block1_conv1 (Conv2D)	(224,224,64)	1792
block1_conv2 (Conv2D)	(224,224,64)	36928
block1_pool (MaxPooling2D)	(112,112,64)	0
block2_conv1 (Conv2D)	(112,112,128)	73856
block2_conv2 (Conv2D)	(112,112,128)	147584
block2_pool (MaxPooling2D)	(56,56,128)	0
block3_conv1 (Conv2D)	(56,56,256)	295168
block3_conv2 (Conv2D)	(56,56,256)	590080
block3_conv3 (Conv2D)	(56,56,256)	590080
block3_pool (MaxPooling2D)	(28,28,256)	0
block4_conv1 (Conv2D)	(28,28,512)	1180160
block4_conv2 (Conv2D)	(28,28,512)	2359808
block4_conv3 (Conv2D)	(28,28,512)	2359808
block4_pool (MaxPooling2D)	(14,14,512)	0
block5_conv1 (Conv2D)	(14,14,512)	2359808
block5_conv2 (Conv2D)	(14,14,512)	2359808
block5_conv3 (Conv2D)	(14,14,512)	2359808
block5_pool (MaxPooling2D)	(7,7,512)	0

flatten (Flatten)	25088	0
dense (Dense)	2	50178
Total params	14764866	
Trainable params	50178	
Non-trainable params	14714688	

Table 7.4-1.a: Layers of the transfer learning VGG16 model

flatten (Flatten)	25088	0
dense (Dense)	2	50178
Total params	14764866	
Trainable params	14764867	
Non-trainable params	0	

Table 7.4-1.b: Layers of the VGG16 model from scratch

As we can notice from the above tables, the architecture of the network is the same, with the only difference in the number of parameters. In fact, in the case of transfer learning (table 7.4-1.a), most of the parameters are trained based on the ImageNet dataset, so only 50178 parameters are fine-tuned with the collected dataset. On the other hand, all parameters are trained from scratch.

7.5 Achieved results

In order to obtain the best possible result, the network was trained varying various parameters, such as the size of the batch used for training and validating the models, the number of training epochs and the optimizer learning rate. This trial and error process required several hours of network training, producing a model with discrete loss function and accuracy values.

The graphs are reported in *Figure 7.5.1.a-b* shows the representative values obtained during the training. More precisely, the saved model corresponds to the model trained in the epoche that corresponds to a lower loss function value, thus avoiding overfitting.

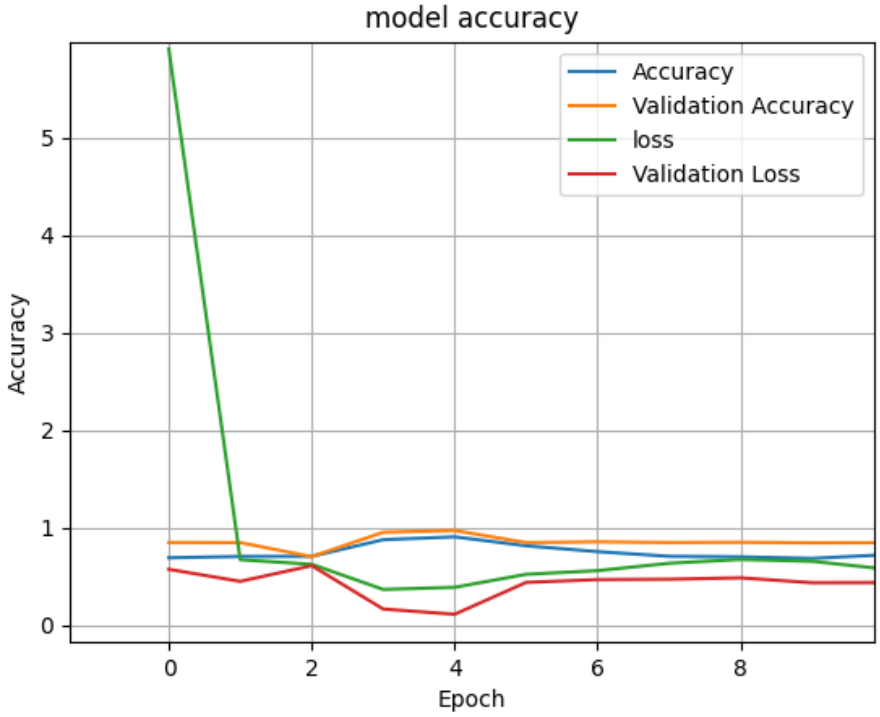


Figure 7.5.1: Results obtained from the re-trained network.

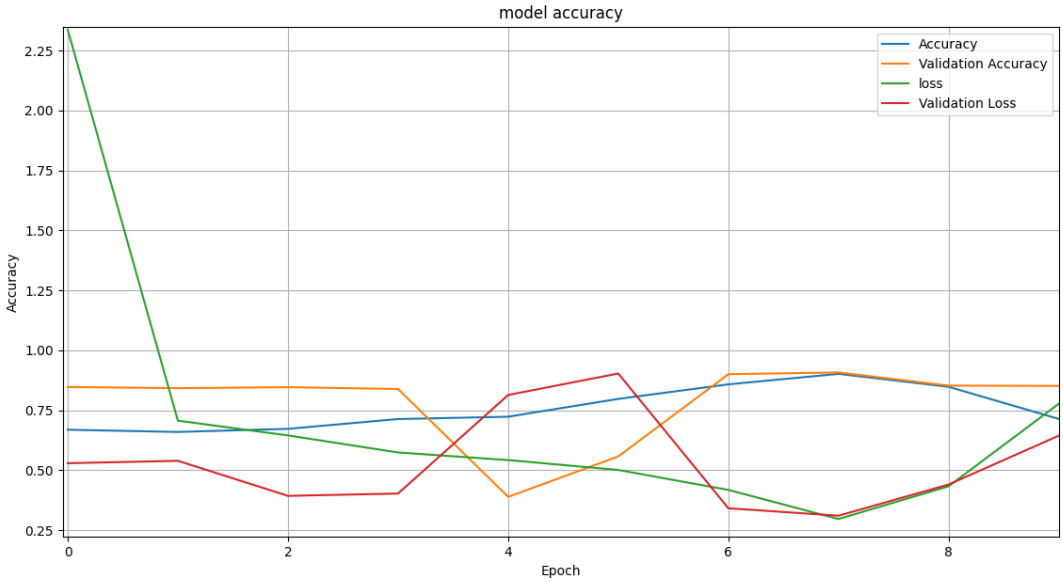


Figure 7.5.2: Result obtained from trained the network from scratch.

As can be seen from the graphs shown, both methods converge to a model with promising accuracy values. in the first case, the optimal model is reached after 3 training periods with accuracy values of approximately 95%. In the second case, however, the optimal model was reached after 7 training periods with a value slightly lower than the previous one, equal to 93%.

Despite the results obtained, and in order to optimize the predictions of the neural network, the model was implemented in a software capable of analyzing several successive frames relating to the audio acquisition.
With this solution we wanted to analyze several audio segments, in order to be sure of the result obtained. In fact, it has been noticed that sometimes the acquisition of a segment related to the audio of a defective component, was not very visible from the neural network, finding false predictions of the sample. In the proposed solution, therefore, more predictions of the same component are taken into consideration, improving the accuracy of the software.

8 Conclusion

In this study, the sound of gearboxes was recorded and processed to classify and identify the defects of the components. For this purpose, an acoustic insolated platform, incorporating a microphone for signal capture was built. The software used was designed with two different approaches. The first, based on signal theory, consists of the audio processing in the time-frequency domain to find possible defects located in the gear of the gearboxes and/or report noises that are out of acceptable standards. Instead, the second approach, implement a convolutional neural network (CNN), which was first re-trained from a network obtained from a dataset called ImageNet and then trained from scratch with an augmented dataset, in such a way to have the comparison between both methods. The result achieved in the two different approaches shows how both are reliable, with the difference that, in the first case, the implemented method works in post-process, as first, the audio is recorded and then classified. However, in the second case, the analysis can take place in real-time, analyzing shorter audio segments. For future projects, it would be interesting to invest more resources so that to implement an accelerometer as a vibration acquisition system and make a cross-comparison between audio and vibration. Furthermore, the development and improvement of the second proposed approach could be implemented in embedded microcontrollers to have the possibility of continuous onboard control.

Bibliography

[1] Engineering_ToolBox. *STP - Standard Temperature and Pressure & NTP - Normal Temperature and Pressure.* 2004. https://www.engineeringtoolbox.com/stp-standard-ntp-normal-air-d_772.html.

[2] Rumsey, F., and T. McCormick. *Sound and Recording: An Introduction.* V. CRC Press, 2006.

[3] Sataloff, R.,T, and J. Sataloff. *Occupational Hearing Loss.* Boca Raton: CRC Press, 2006.

[4] Acoustical Society of America. "Welcome to ASA Standards." 14 December 2020.

[5] Bilinskis, Ivars. *Digital Alias-free Signal Processing.* John Wiley & Sons, 2007.

[6] Velardo, V. "Audio signal processing for machine learning." 2020.

[7] Gentenbein, M. *Microphone Basics – Part 2: Polar Patterns – Mike's Media Production Tips.* n.d. http://www.mikegantenbein.com/blogs/uncategorized/microphone-basics-part-2-polar-patterns/.

[8] Delta Electronics, Inc. *A High Performance with Diverse Communication INterfaces Servo Drive ASDA-A2R Series User manual.* 2014.

[9] Modbus Organization. "MODBUS over Serial Line Specification and Implementation Guide V1.02." *modbus.org.* 20 DEC 2006.

[10] Bertazioli, O. *Tecniche di modulazione analogiche.* in riga edizioni, 2015.

[11] Stremler, F. *Introduction to Communications Systems.* third edition. Addison Wesley Publishing Company, 1992.

[12] HP application Note. "Spectrum Analysis Basic." n.d.

[13] Weisstein, Eric W. *Autocorrelation.* n.d. https://mathworld.wolfram.com/Autocorrelation.html.

[14] McCorduck, P. "Machines Who Think." In *Machines Who Think, A Personal Inquiry into the History and Prospects of Artificial Intelligence*, 598. A K Peters/CRC Press, 2004.

[15] Stuart J.; Norvig P. "Artificial Intelligence a Modern Approach." Englewood Cliffs, New Jersey: Prentice Hall, n.d.

[16] Marvin Minsky, Seymour Papert. "Perceptrons: an introduction to computational geometry." Massachusetts, 1969.

[17] Abraham, Ajith. "Artificial Neural Networks." In *Handbook of measuring system design* . 2005.

[18] Haykin, Simon. *Neural Networks and Learning Machines.* Third edit. Ontario: Pearson Education, 2009.

[19] Albawi S, Bayat O, Al-Azawi S, Ucan ON. "Social Touch Gesture Recognition Using Convolutional Neural Network." In *Computational intelligence and neuroscience*. 2018.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." *IJCV.* 2015.