UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI INGEGNERIA

Corso di Laurea in Ingegneria Meccanica

Analysis of Sound quality metrics and their correlation with subjective data using neural network algorithms

Analisi delle metriche del Sound quality e loro correlazione con i dati soggettivi mediante algoritmi di reti neurali

Relatore:
Prof. Paolo Castellini

Tesi di:
Mattia Elisei

Anno accademico 2020/2021

Table of contents

List of figures

Abstract

The utilization of modern computers' computational capacity has become a decisive element in addressing engineering issues. The advancement of neural network algorithms, in particular, has enabled these computers to "learn" and "recognize," tasks that were previously exclusive to the human mind.

These algorithms may be used to evaluate sound sources and determine which ones will irritate the human ear in a sound quality research. This is only feasible if the neural networks have been adequately "trained" on a huge database of noises, giving them a sense of how annoying each one is.

It is easy to believe that there is a link between the listener's irritation and certain aspects of the sound they are hearing. The metrics define the major features of a sound.

The objective of this thesis is to use neural network algorithms to evaluate sound metrics and establish a link with subjective data.

Abstract: italian version

Negli ultimi decenni, l'utilizzo della potenza di calcolo dei nostri computer è diventato un fattore determinante per la risoluzione di problemi ingegneristici. In particolare, lo sviluppo di algoritmi di reti neurali ha permesso a queste macchine di “apprendere” e “riconoscere”, funzioni che, fino a poco tempo fa, erano relegate alla sola mente umana.

Nello studio della qualità del suono, questi algoritmi possono essere utilizzati per analizzare sorgenti sonore ed identificare quali risulteranno fastidiose all'orecchio umano. Questo è possibile solo se le reti neurali sono opportunamente “allenate” su un grande database di suoni, fornendo loro un'indicazione della fastidiosità di ognuno di essi.

È naturale pensare che esista una relazione tra la fastidiosità percepita dall'ascoltatore e alcune caratteristiche del suono ascoltato. Le principali caratteristiche di un suono sono descritte dalle metriche.

L'obiettivo di questa tesi è quello di analizzare le metriche del suono e di trovare una correlazione con i dati soggettivi utilizzando algoritmi di reti neurali.

Ringraziamenti

Vorrei ringraziare il mio relatore, il professor Paolo Castellini, per avermi permesso di svolgere il tirocinio e la tesi nel laboratorio di vibrazioni dell'Università Politecnica delle Marche.

È stata un'esperienza formativa sia dal punto di vista delle conoscenze apprese, sia dal punto di vista delle relazioni e delle interazioni avute nel laboratorio con il professore e i suoi assistenti, aspetto non secondario visto l'ultimo anno svolto quasi completamente in didattica a distanza a causa della pandemia di COVID – 19.

In modo particolare intendo ringraziare Reza Jamali, l'assistente che ha maggiormente contribuito al lavoro svolto durante il tirocinio e quindi anche alla produzione di questa tesi.

Infine ringrazio la mia famiglia che mi ha continuamente supportato durante questi mesi di continue restrizioni e limitazioni.

1 – Introduction

A study was done into how a number of sound quality metrics may be used to anticipate consumer reactions to noises from a certain type of product, as represented in product-specific characteristics such as ratings of "acceptability" and "quality." We presume that a jury study has previously been done for such qualities, resulting in rating values for various product sounds, with the goal of determining which metrics or combinations of metrics may best be utilized to forecast user judgements for different versions of the product's sounds.

In this thesis, we focused on the study of sounds from kitchen hoods and washing machines.

Considering the hoods, we collected the judgements made by questioned listeners on the perceived unpleasant sound, in addition to the audio files.

We computed the major sound metrics in Matlab and linked them to the respondents' comments using neural network techniques. We utilized and evaluated several calculating softwares for the actual calculation of the metrics: Matlab, Testlab, and Bruel.

Finding this link would entail anticipating the sensations a person has while listening to a new sound with a high degree of accuracy. This could only be accomplished by analyzing the signal rather than interviewing individuals, saving both time and money.

2 – Metrics

2.1 – Loudness

The loudness of a sound is measured in sones. It's not the same as decibels. Sones account for the frequency and level dependent aspect of human hearing, whereas decibels do not fully account for this dependency.

Many individuals believe that utilizing a decibel number to quantify the volume of a sound is the best way to do it. The volume of a sound is properly represented by decibel (dB) values, however the perceived loudness of a sound is not accurately represented by decibel (dB). In reality, there is a different sound quality metric called loudness (with units of sones or phons) that represents how humans perceive the volume of a sound considerably better.

Figure 1 illustrates the human hearing domain. When looking at the lower limit (hearing threshold), it is clear that it fluctuates with frequency. At different frequencies, the threshold has various values.

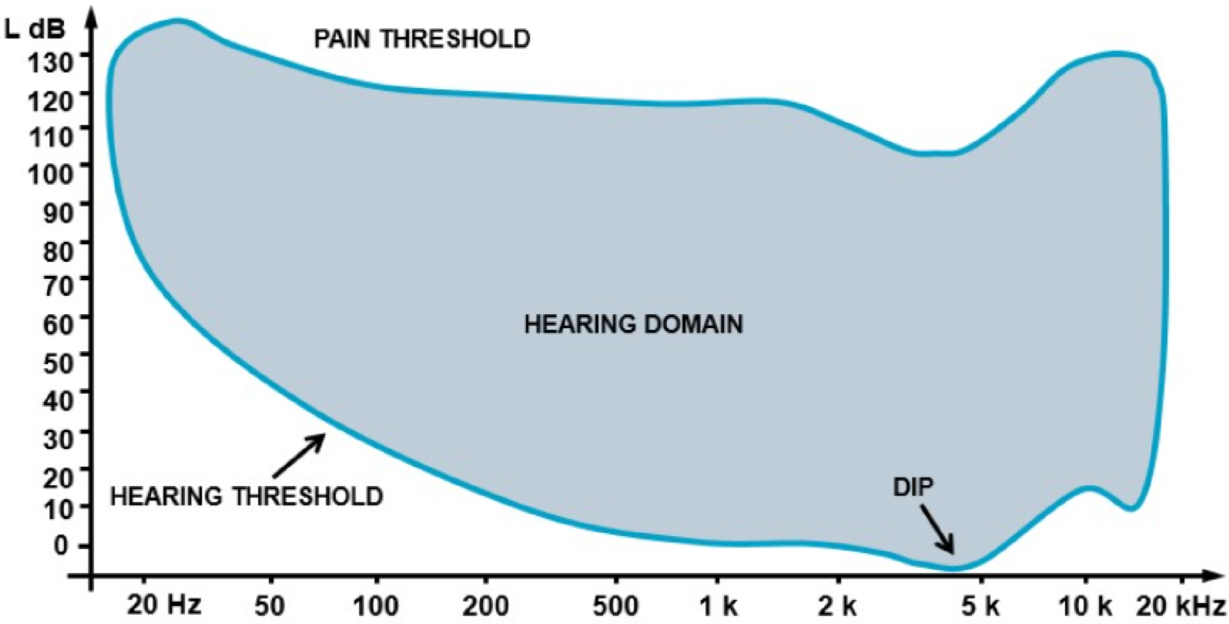

Figure 1: The human hearing domain is both frequency and level dependent. The dip at about 4000Hz is because the air volume in the ear canal goes into resonance at that frequency.

Figure 1 shows a dip between 3000Hz and 5000Hz. At these frequencies, humans can hear very well. Between 3 and 5 kHz, people can hear sounds at lower decibel levels than at any other frequency.

A 10dB sound at 5 kHz, for example, is audible to humans, while a 10dB sound at 50Hz is not.

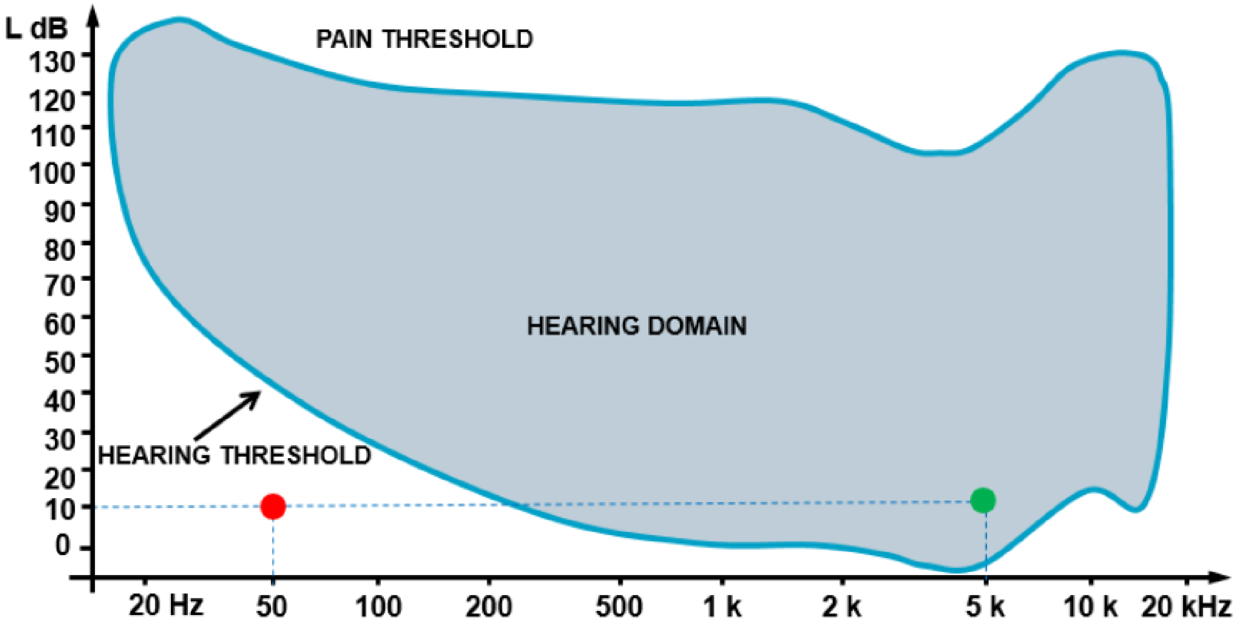

Figure 2: A tone at 50Hz and 10dB is inaudible. A tone at 5000Hz and 10dB is audible.

Both frequencies have the same dB value, yet the perceived loudness is vastly different — one is audible, while the other is inaudible. Clearly, dB is insufficient to describe a sound's perceived loudness.

The loudness measure is based on how loud something is judged to be. As a result, the measure was created in collaboration with a human jury (unlike decibels which is simply a math equation). Each curve in the graph below shows a curve for sinusoidal tones of equal loudness. To produce an equally loud sound, the dB value must vary as frequency varies along a curve.

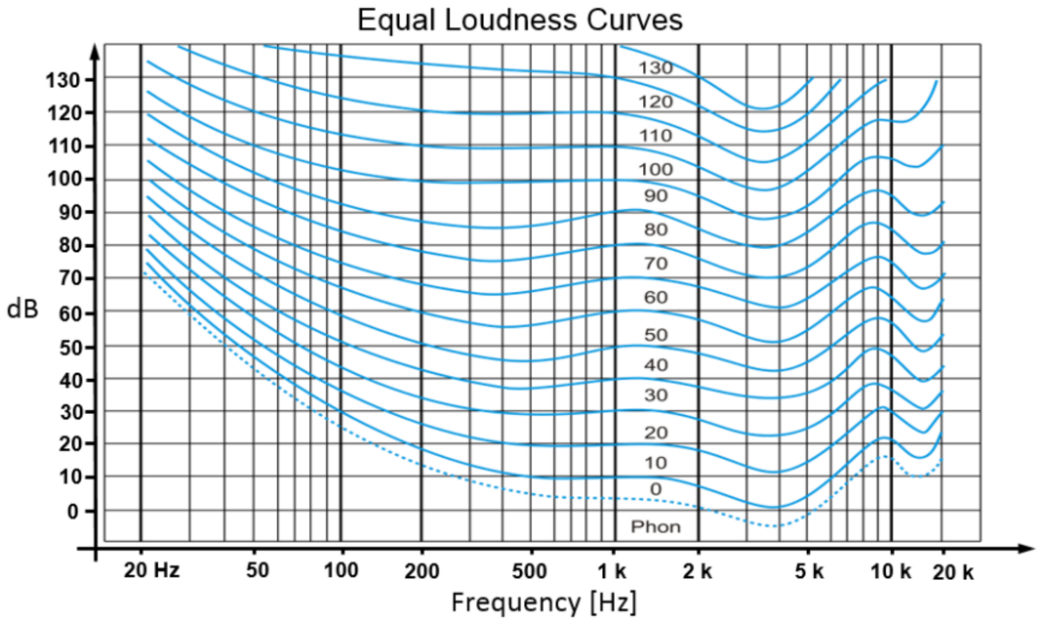

Figure 3: The curves of equal loudness originally developed by Fletcher-Munson in 1933.

A jury of individuals with normal hearing was assembled to develop this measure. The jurors would hear a tone at 1000Hz with a specific dB level. Then, at a different frequency, a second tone would be played. The second tone's volume would be adjusted until it sounded as loud as the 1000Hz tone.

The loudness level is measured in sones or phons, which are both loudness units. Because the sone is a linear unit, it is usually chosen over the phon.

2.2 – Sharpness

Sharpness is a metric that measures the balance of a sound's spectral richness between low and high frequencies. If a spectrum's energy is mainly concentrated in the low frequency band, the sharpness value will be low. The sharpness value will be high if the signal's energy is skewed toward the high frequency end of the hearing spectrum. An intermediate sharpness rating is produced by a flat spectrum with well-balanced energy across the frequency range.

Sharpness, like Loudness, Tonality, and other sound quality indicators, may be used to distinguish between sounds that may have the same total decibel level but generate quite distinct subjective impressions.

The unit of acum, which is derived from the Latin word for "sharp," is used to measure sharpness. A narrowband noise with a critical band width of one kHz and a level of 60 dB RMS is defined as 1.0 acum. The numbers generated by the sharpness computation are always non-negative integers, and their value is potentially limitless (depending on the formulation used, more on that in the next section). The sharpness of a signal with no content in audible frequencies is 0.0 acum.

The specific loudness spectrum, or loudness versus the Bark scale, is the starting point for calculating sharpness. The overall loudness may be determined using the specific loudness spectrum and then utilized in the sharpness calculation. Take two well-known test signals, White and Pink Noise, to demonstrate this procedure. Figure 4 illustrates the narrow band frequency spectra of these two signals.

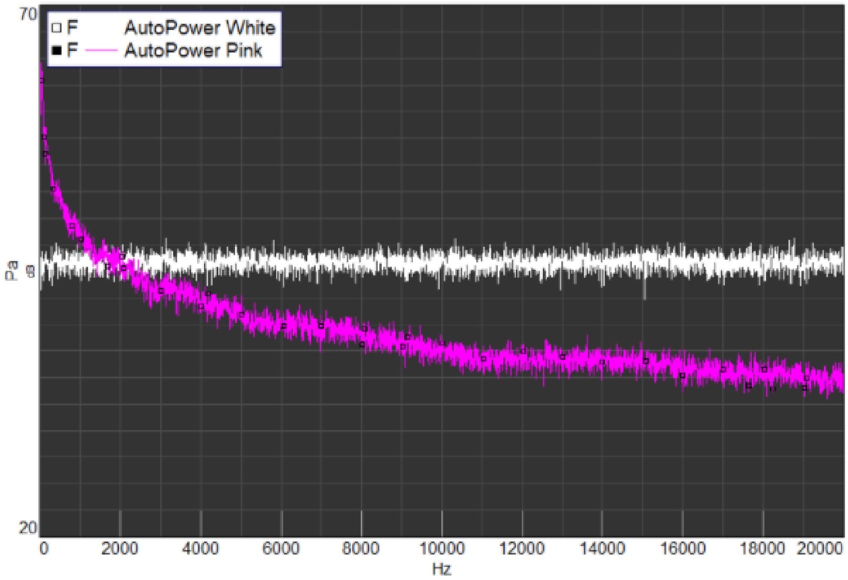

Figure 4: Narrow-band frequency spectra for white and pink noise.

As demonstrated by the flat white line in Figure 4, white noise is defined by equal energy across the narrow-band frequency spectrum, or equal amplitude across all frequencies. The amplitude level of pink noise drops off continuously as a function of narrow-band frequency, as shown by the pink trace in Figure 4. Pink noise is energy-biased toward low frequency, and the amplitude level drops off continuously as a function of narrow-band frequency, as shown by the pink trace in Figure 4.

Calculating the specific loudness spectrum for each sound is the first step in calculating sharpness. The psycho-acoustic loudness of a signal is represented by the specific loudness spectrum, which is spread throughout the frequency bands that people perceive in, known as Bark bands or the Bark scale.

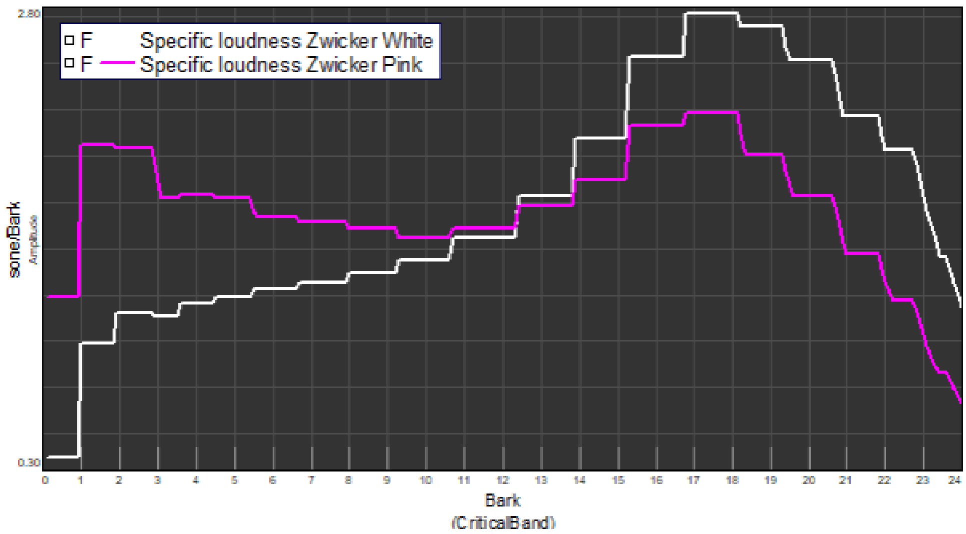

Figure 5: Specific loudness spectra for white and pink noise. Specific loudness specifies the distribution of loudness over the 24 Bark bands.

The loudness of pink noise is generally balanced between low and high Bark bands, whereas white noise is skewed toward the higher Bark bands, as seen in Figure 5.

The total loudness (N), which is determined from the specific loudness spectrum, is the next step in the sharpness computation. In Figure 6, the equation for overall loudness is displayed. Total loudness is calculated by integrating the specific loudness spectrum over the 24 Bark bands and converting it to Sones.

$$N = \int_0^{24} N'(z) \cdot dz$$

Figure 6: Formulation of total loudness (N) calculated from specific loudness spectrum (N').

Sharpness may be calculated using a number of different formulas and criteria.

Zwicker sharpness, for example, is determined using the method given in Figure 7. Sharpness is defined as a ratio of the weighted total loudness of the spectrum to the overall loudness, as shown in the equation.

$$S = k \frac{\boxed{\int_0^{24} N'(z)} \cdot g(z) \cdot z/Bark \cdot dz}{\boxed{\int_0^{24} N'(z)} \cdot dz}$$

Total Loudness

Figure 7: Formulation of Sharpness (S) for Zwicker and DIN45692 methods.

2.3 – Fluctuation Strenght and Roughness

These two sound metrics, unlike decibels, which only measures the absolute intensity of sound, combine the following characteristics of a sound (Figure 1) into a single number:

- **Modulation Frequency (fmod)** – The number of peaks and falls in the sound per second
- **Modulation Level (ΔL)** – The perceived magnitude level change throughout time

The greater the values of Fluctuation Strength or Roughness, the more visible the modulation.

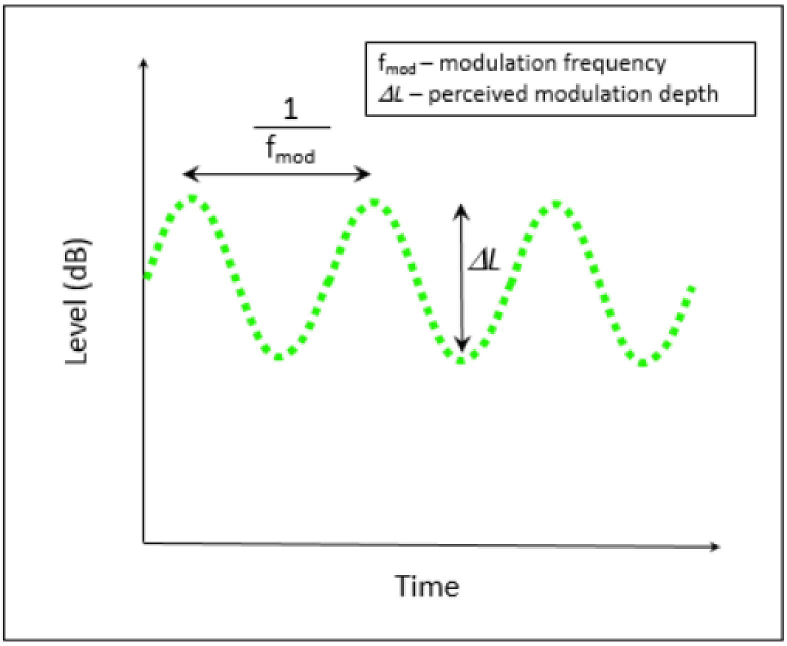

Figure 8: The amplitude modulation of a sound is described by a frequency and level

The Fluctuation Strength or Roughness measure may be suitable depending on the amount of modulations per second present in the sound:

- **Fluctuation Strength**: Describes sounds with less than 20 modulations per second.
- **Roughness**: Describes sounds with a minimum of 20 modulations per second and a maximum of 300

Fluctuation Strength may be used to measure low-frequency modulations such as propeller plane droning, exhaust rumbling, or the lugging of an electric motor.

Roughness may be used to measure high-frequency modulations such as an electronic razor's buzzing, a fan's fast blade passing noise, or fuel injectors' "sewing machine noise."

There are two primary explanations for the fluctuation in sound levels over time:

1. Amplitude: The sound's amplitude, or level, may rise and fall over time, as if someone were turning the radio's volume knob up and down. Even if the signal's frequency content remains constant, this can happen.

2. Frequency: The modulation is caused by constructive and destructive interference between several frequency tones present in the sound.

When two tones of equal amplitude but different frequency are played at the same time, this is known as frequency based modulation. The phase between the two tones varies over time due to the frequency difference, as illustrated in Figure 3. The tones are sometimes in phase, and sometimes they are out of phase.

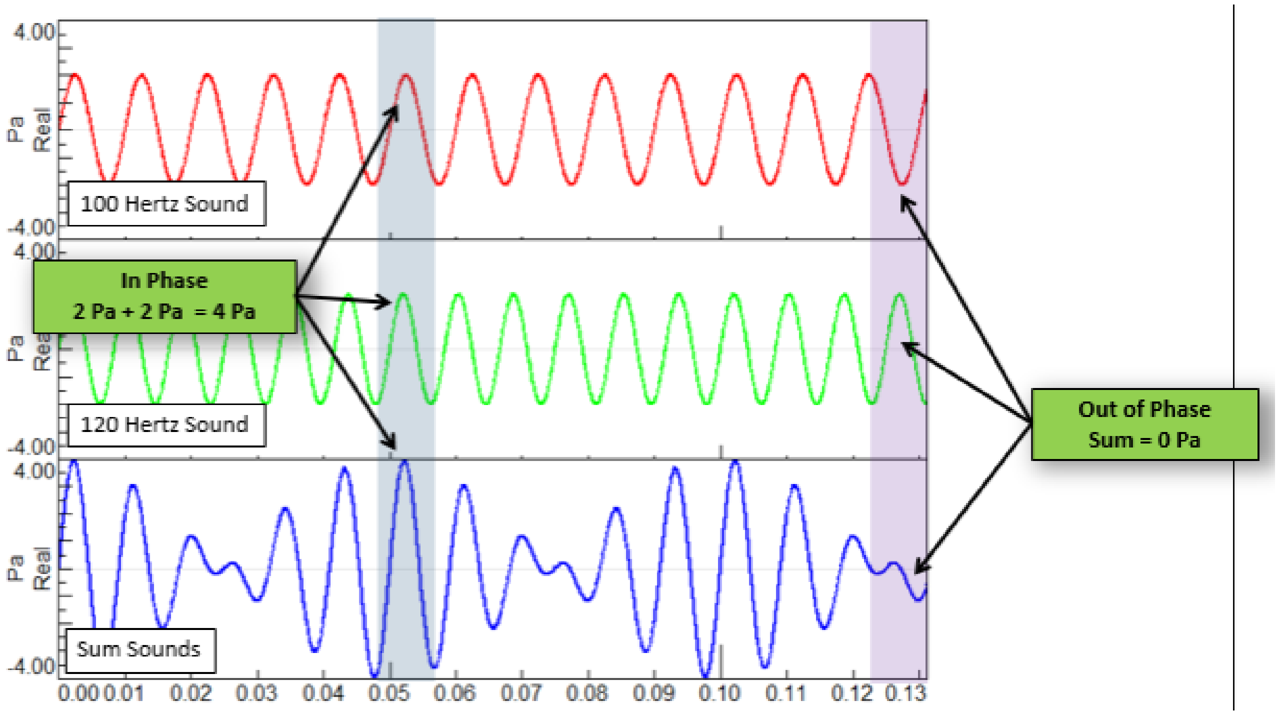

Figure 9: When listened to simultaneously, a 100 Hertz tone and 120 Hertz tone will constructively and destructively interfere with each other

Modulation frequencies below 20 Hz are described by Fluctuation Strength. A listener can hear each individual rise and fall in the sound because the sound fluctuates slowly over time (below 20 modulations per second).

Fluctuation Strength is measured in Vacil units. This is a short version of the English word vacillate, and originates from the Latin word vacillātus. It's worth noting that the vacil value stabilizes after at least one modulation, which is an essential factor when computing this metric. It is critical to use the metric values from the stabilized section for a steady state signal.

Modulations that occur more than 20 times per second and up to 300 times per second are described as rough. The human ear is unable to identify individual modulations when they occur more than 20 times per second.

There is a sense of a steady, yet harsh tone with modulations of 20 to 150 times per second. Listeners frequently claim hearing three distinct tones at higher modulation frequencies (about 150 to 300 times per second). The "asper" is the unit used to indicate roughness.

2.4 – Articulation Index

The Articulation Index (AI) is a sound metric that measures how much background noise interferes with human speech.

Its value ranges from 0% (no speech comprehended) to 100% (complete understanding) (all speech understood).

Initially, the Articulation Index was designed to assess voice privacy and communication system performance. Articulation Index is now used to assess car interior noise, white goods quietness, and other factors.

Humans can hear frequencies ranging from 20 to 20,000 Hertz, but the frequencies generated in human speech are significantly smaller. The frequency and amplitudes that a person can hear are highlighted in light blue in Figure 10. The frequencies and amplitudes produced by typical human speech are shown in orange.

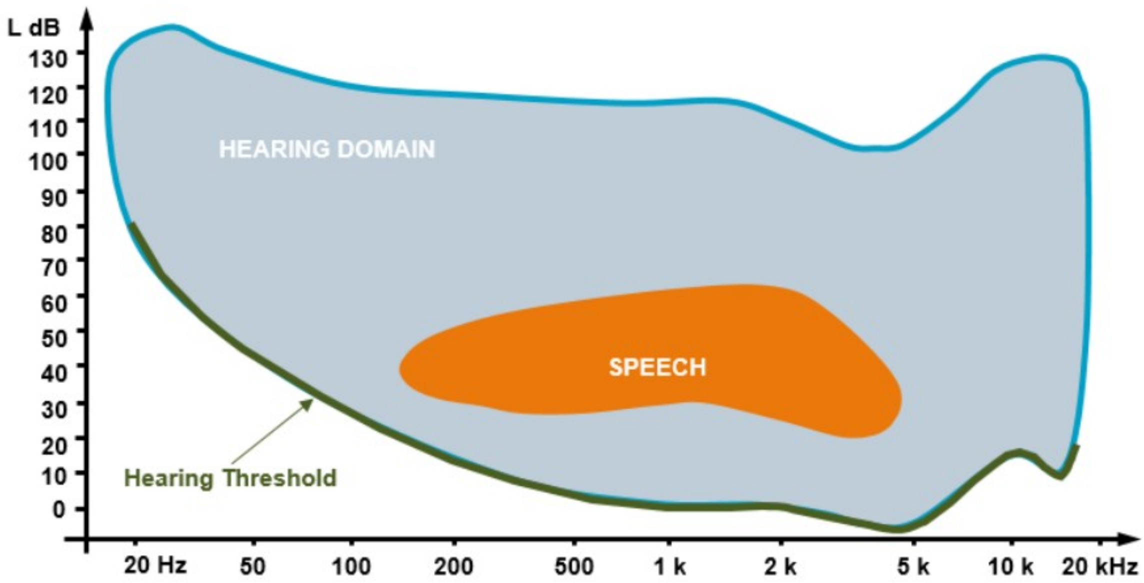

Figure 10: Map of human hearing audio range with sound level in decibels versus frequency (light blue). The human speech frequency range (orange), is critical to speech being understood properly.

The speech frequency ranges from 200 to 6000 Hertz. The background sound levels that occur within this frequency range are given the highest priority when calculating the Articulation Index, as they will interfere with human speaking. Outside of this range, background noises are unimportant.

During World War II, Leo Beranek of Harvard University created one of the earliest definitions of Articulation Index. He utilized it to assess the efficacy of several aircraft headsets. While each standard may differ somewhat, the following procedures are usually used for calculating Articulation Index:

1) Perform a background sound measurement at the place where a listener is positioned.

For instance, the interior of a vehicle at highway speeds, the interior of a cockpit during a flight, and so forth. Broadband background noises are commonly utilized with Articulation Index.

2) From the measurement in the preceding step, calculate an A-weighted octave spectrum in decibels.

3) Plot the background sound against the chart in Figure 11 to see how much each 1/3rd octave band's articulation window is "covered."

When the sound level exceeds the articulation upper limit (red line in Figure 11), the octave band's window is entirely "covered." The octave band is partially "covered" if the level lies between the top and lower boundaries.

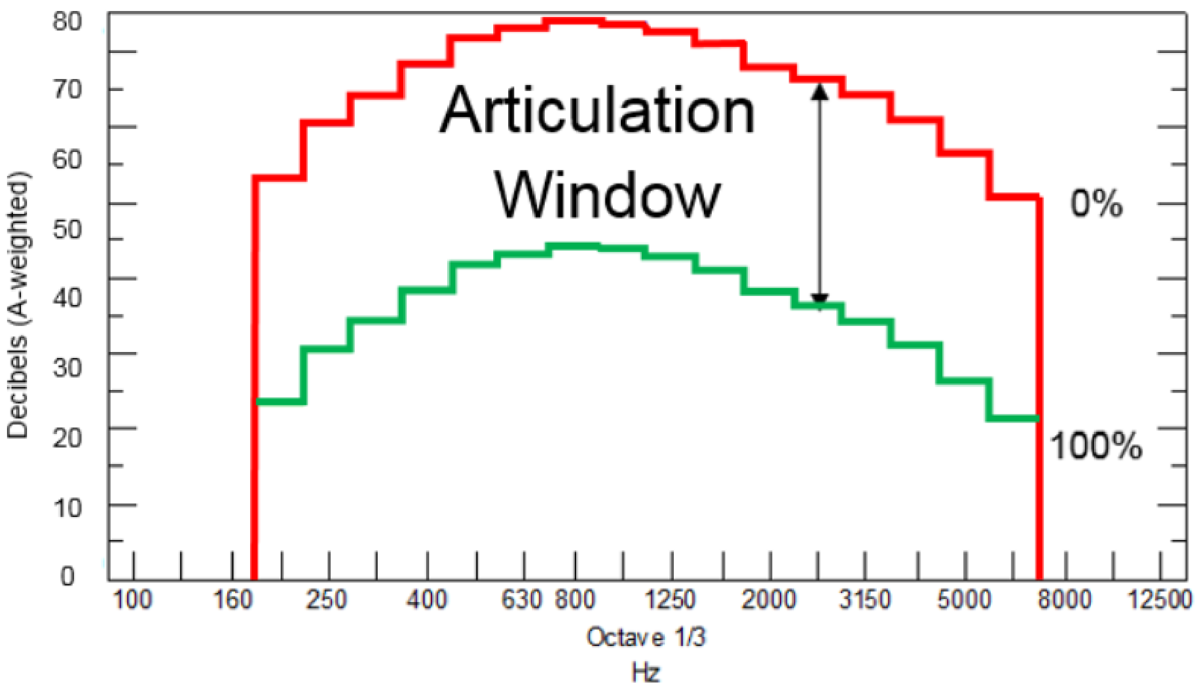

Figure 11: The Articulation window is defined by a lower and upper limit, separated by 30 dB, for each octave band.

For each 1/3 octave band, a coverage value is determined.

Figure 12 shows the exact numbers that define the top and lower boundaries of the articulation window. The coverage values are then weighted in the following phase of the computation.

The weighting variables mentioned in Figure 12 are multiplied by the coverage values from the previous step.

1/3rd Octave Center Frequency (Hz)	AI lower level dB(A)	AI upper level dB(A)	Weighting Factor
200	23.1	53.1	1
250	30.4	60.4	2
315	34.4	64.4	3.25
400	38.2	68.2	4.25
500	41.8	71.8	4.5
630	43.1	73.1	5.25
800	44.2	74.2	6.5
1000	44.0	74.0	7.25
1250	42.6	72.6	8.5
1600	41.0	71.0	11.5
2000	38.2	68.2	11.0
2500	36.3	66.3	9.5
3150	34.2	64.2	9.0
4000	31.0	61.0	7.75
5000	26.5	56.5	6.25
6300	20.9	50.9	2.5

Figure 12: Articulation Index weighting factors for octaves from 200 Hz to 6300 Hz.

The weighted coverage values for each octave band are totaled after weighting to generate a single Articulation Index number.

3 - Neural networks

3.1 – Introduction to neural networks

Today, most of the technology we use on a daily basis is based on a sophisticated structure of algorithms that are based on neural network theory. In actuality, these "things" are nothing more than more or less complicated algorithms that may "learn" to generalize specific ideas based on a training data set. There are several forms of neural networks that, depending on the situation, can be utilized to achieve the goals we have set for ourselves.

A neural network is a mathematical model with layers of linked nodes that resembles the layered structure of the brain's network of neurons. A neural network may be trained to identify patterns, categorize data, and predict future events by learning from data.

The biochemistry of a live being's brain is extremely complex: the nervous system has several billion neurons. The first, the dendrites, are branching extensions via which they receive electrical signals from other neurons (inputs); the second, the axons, are extensions of various lengths (from 1cm to a few meters) that finish in branches and serve to convey an electrical signal to other cells.

In order to artificially replicate a human brain, we need a network of simple elements with a large distribution and the ability to function in parallel in order to learn and generalize (which means creating new knowledge starting from the basic elements learned in training). An artificial neural network's primary structure is the artificial neuron, which is a "thing" with multiple inputs and a single output. Each input has its own weight, which is the value of the input signal's intensity (conductivity). The weighted total of the inputs causes the neuron to turn on.

3.2 – Theory

The inputs to a neural network are broken down into layers of abstraction. It, like the human brain, can be trained to identify speech or picture patterns using a variety of examples. The way its component pieces are linked, as well as the strength, or weights, of those connections, determine its behavior. These weights are changed automatically during training according to a set of rules until the neural network completes the job properly.

A neural network is a type of technology that integrates multiple levels of processing by employing basic pieces that run in parallel and is inspired by organic nerve systems. An input layer, one or more hidden layers, and an output layer make up the structure. Layers are connected by nodes or neurons, with each layer taking the preceding layer's output as input.

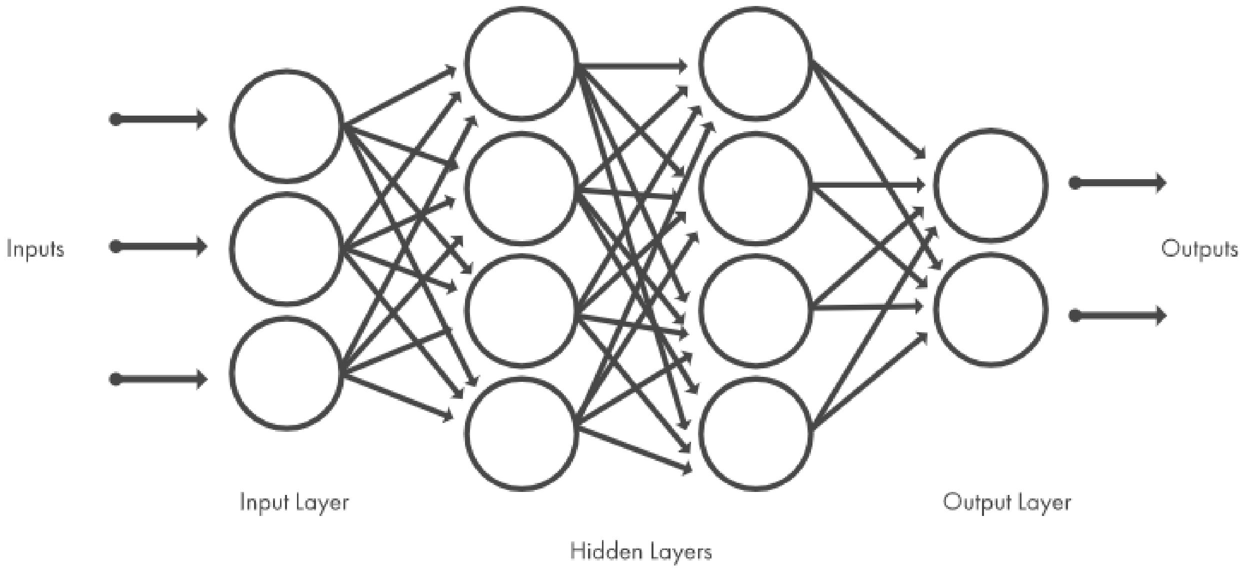

Figure 13: Typical neural network architecture.

The following are some of the most common machine learning approaches for developing neural network applications:

-**Supervised learning**: supervised neural networks are taught to generate desired outputs in response to sample inputs, making them especially well-suited to modeling and managing dynamic systems, categorizing noisy data, and forecasting future occurrences.

-**Classification**: a form of supervised machine learning in which an algorithm "learns" to categorize fresh observations from examples of labeled data.

-**Regression**: the connection between a response variable (output) and one or more explanatory factors is described by regression models (input).

-**Pattern recognition**: in computer vision, radar processing, speech recognition, and text classification, pattern recognition is a critical component of neural network applications. It works by utilizing supervised or unsupervised categorization to divide incoming data into objects or classes based on essential features.

-**Uns supervision learning**: the neural network is trained in an unsusserved manner by allowing it to adapt to new inputs on a continual basis. They are used to make inferences from datasets that do not have any labeled answers.

-**Clustering**: clustering is a non-obvious learning method that use neural networks for exploratory data analysis to uncover hidden patterns or groups in data. This procedure entails categorizing data based on their similarity.

3.3 - Typical neural network design methodology

Although each neural network application is different, the following stages are usually followed when developing a network:

1) Get data and prepare it

2) Build a neural network

3) Set up the network's inputs and outputs

4) Optimize network parameters (weights and biases) for best results.

5) Train your network

6) Check the results of the network

7) Make the network function as part of a production system.

4 – Data analysis

4.1 – Introduction to the Matlab analysis

Our sound quality study is based on a dataset from kitchen hoods and washing machines, as stated in the introduction.

Essentially, our Matlab-based program does two tasks: it produces sound quality metrics for the data set in question and then compares the matrix of metrics with the matrix of interviewees' judgments on sound discomfort.

4.2 – Metrics calculation

For the actual calculation of the sound metrics, we used the functions present in the Matlab Audio Toolbox package.

As a result, we get a sequence of vectors corresponding to the various metrics (loudness, sharpness, roughness and articulation index). The matrix of metrics was then created by combining these vectors.

4.3 – Linear correlations with the interviewees' judgments

When we compare the opinions of the interviewees with the various acoustic metrics, we find a correlation with loudness and fluctuation strenght (as shown in Figure 14 and in Figure 15).

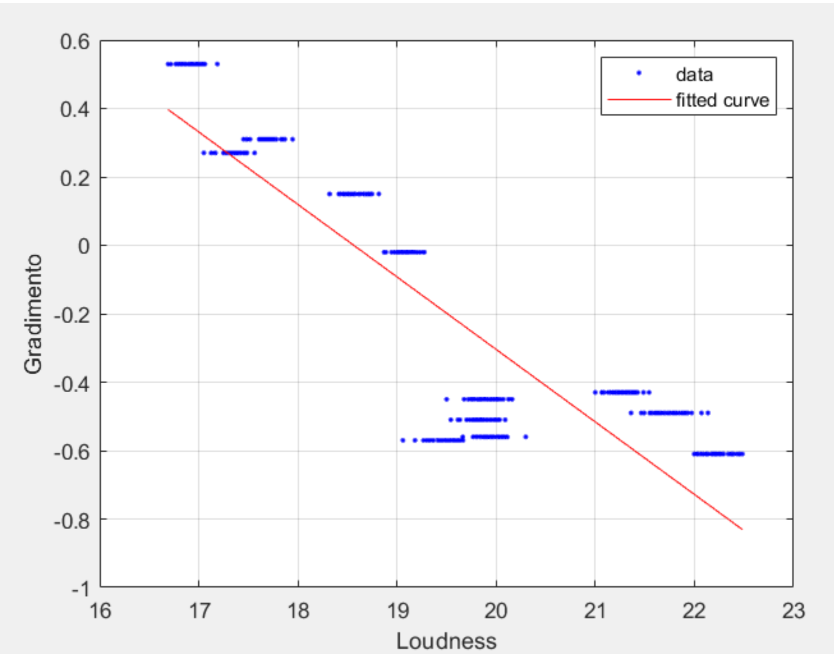

Figure 14: Correlation between respondents' judgment and loudness

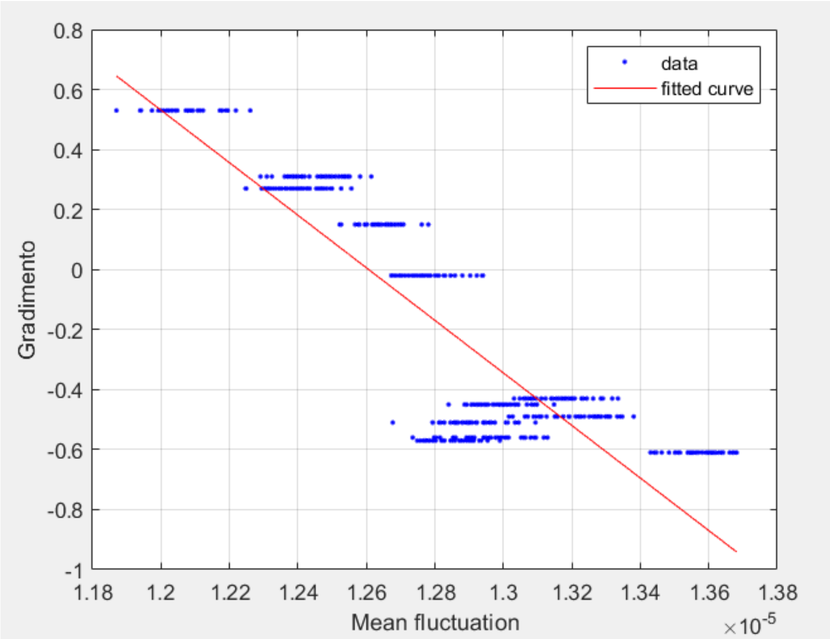

Figure 15: : Correlation between respondents' judgment and fluctuation strenght

As a result, it is clear that the interviewees' judgment deteriorates as the loudness or fluctuation strength increases.

4.4 – Neural network analysis

In this type of analysis we used the Levenberg-Marquardt algorithm to correlate, through neural networks, the matrix of sound metrics with the matrix of judgments.

First and foremost, both matrices must be standardized. The features will be rescaled as a consequence of standardization (or Z-score normalization) to guarantee that the mean and standard deviation are 0 and 1, respectively. The following is the equation (Figure 16):

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

Figure 16: Standardization

This method of rescaling features with a distribution value between 0 and 1 is helpful for optimization methods like gradient descent, which are used in machine learning techniques to weight inputs (e.g., regression and neural networks).

At this point, using the algorithm:

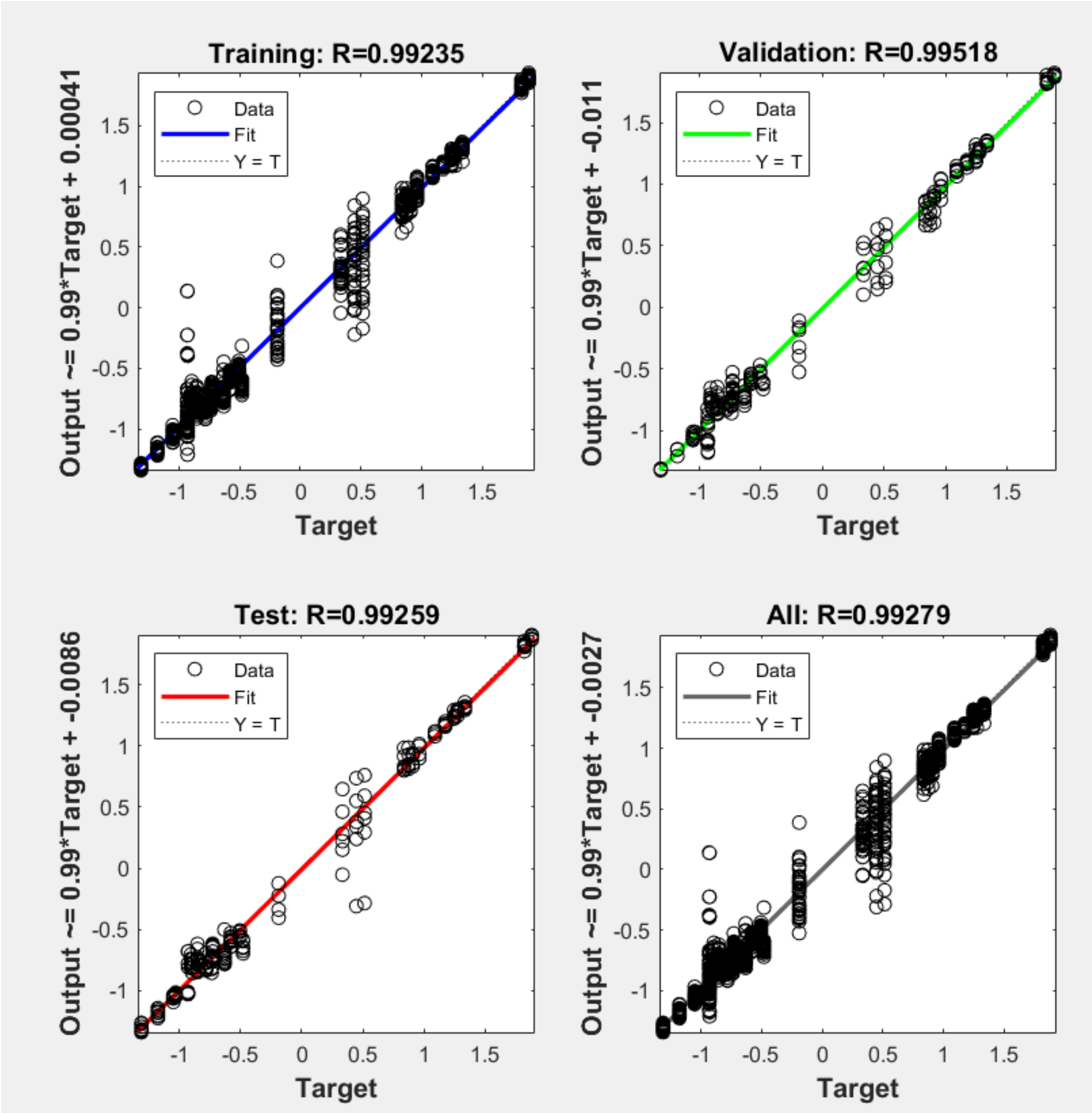

Figure 17: Regression plot

As a result, the input, the matrix of measurements, and the output, the respondents' judgments, have a nearly linear connection. Both matrices had been standardized earlier.

A larger dataset would undoubtedly be required to fully explore the potential of neural networks.

4.5 – Comparison of kitchen hood data supplied by Matlab, Testlab, and Brüel

The metrics study was carried out using three different calculating tools, as previously indicated, so it is fascinating to compare the findings. The graphs below present a comparison of kitchen hood metrics:

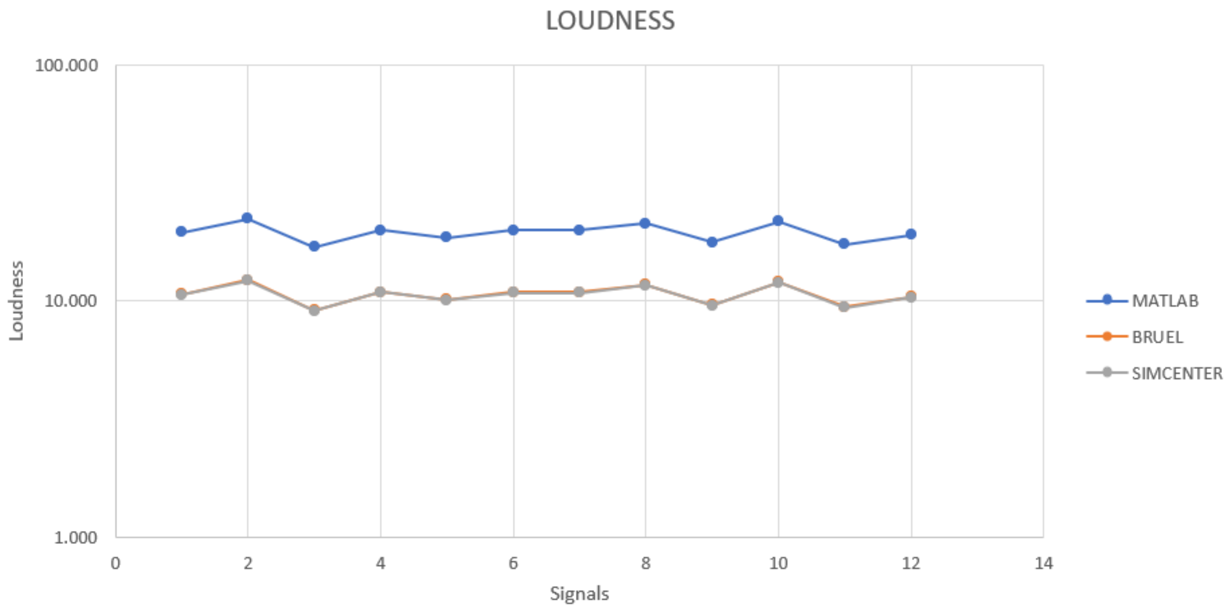

Figure 18: Loudness comparision between Matlab, Testlab and Brüel

Figure 18 shows an almost perfect match between the loudness data supplied by Testlab and those provided by Brüel, but the Matlab data deviate substantially despite retaining the same pattern.

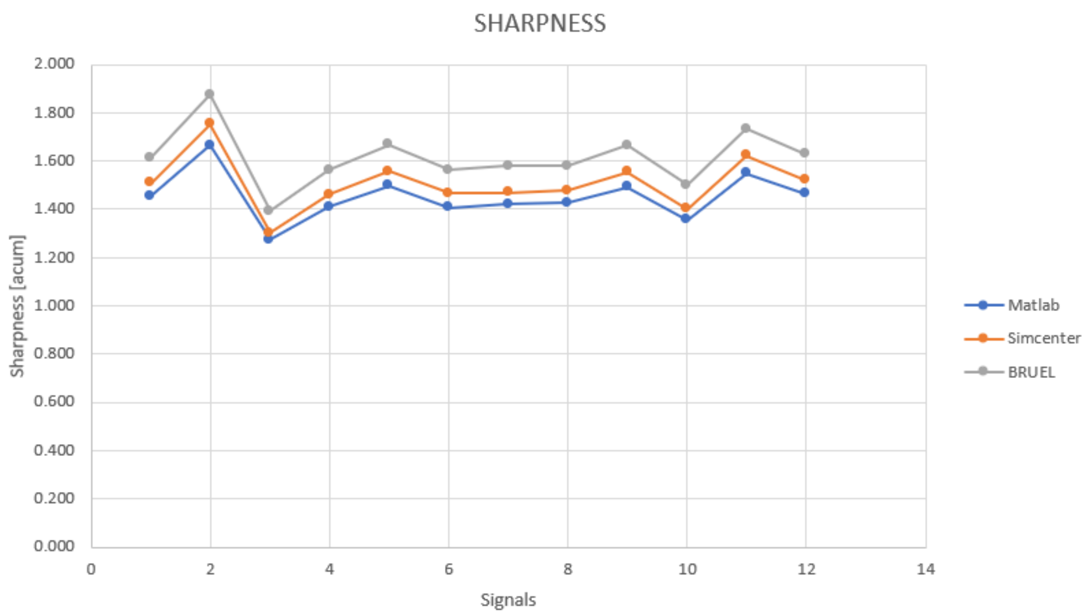

Figure 19: Sharpness comparision between Matlab, Testlab and Brüel

Figure 19 shows a small difference between the sharpness data supplied by the three software programs. The curves look to be misaligned by a considerable offset, but the trends appear to be congruent.

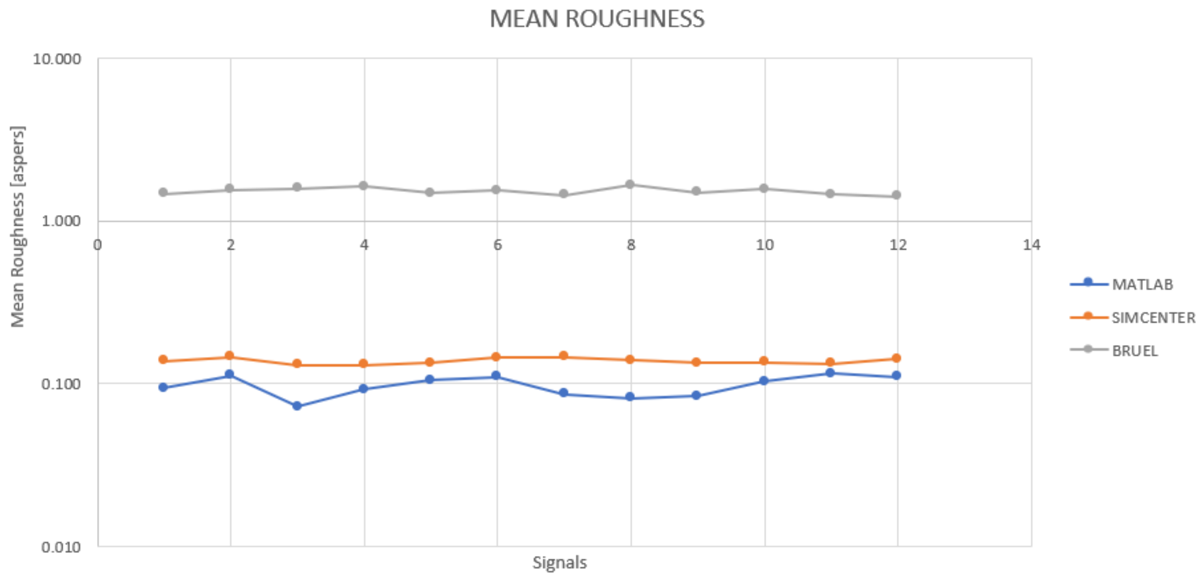

Figure 20: Roughness comparision between Matlab, Testlab and Brüel

Figure 20 shows how the three curves pertaining to the roughness data have three distinct tendencies. The statistics supplied by Testlab and Matlab, in particular, are of the same order of magnitude, but Bruel's are off the charts.

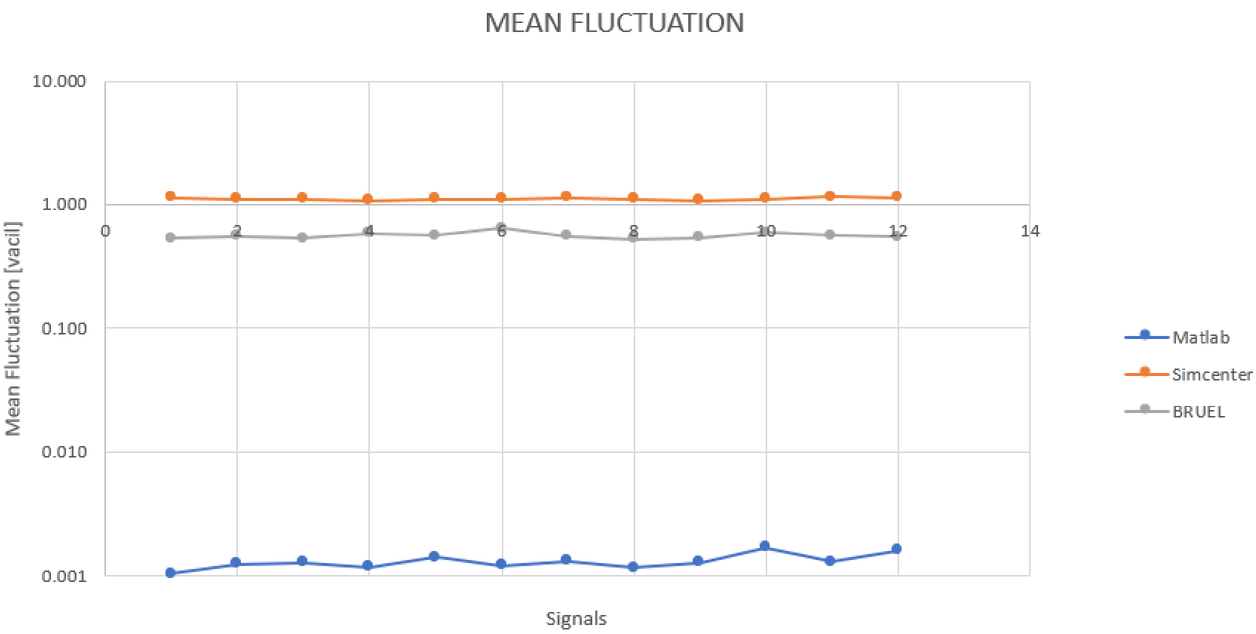

Figure 21: Fluctuation Strenght comparision between Matlab, Testlab and Brüel

Figure 21 shows how the three curves pertaining to fluctuation strength each have their own order of magnitude that differs from the others. They also reveal a variety of trends.

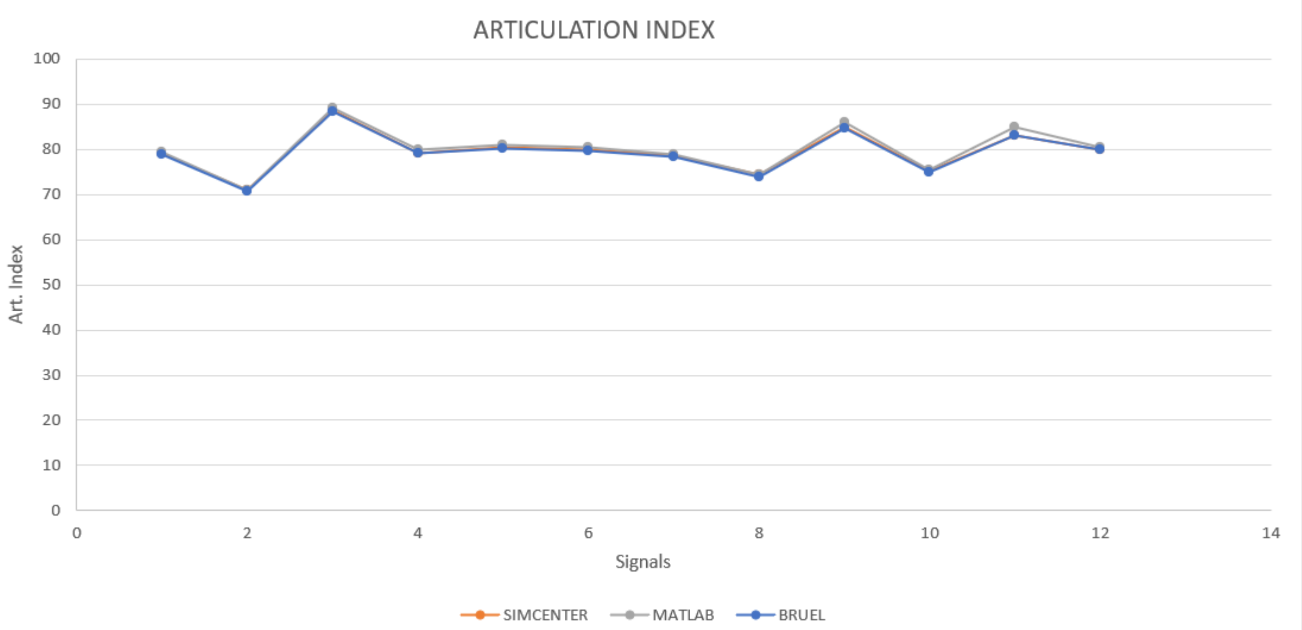

Figure 22: Articulation index comparision between Matlab, Testlab and Brüel

Figure 22 shows an almost perfect correspondence between the data pertaining to the articulation index supplied by the three different computation tools.

The results for the articulation index, loudness, and sharpness can clearly be deemed credible because they are nearly same across the three different calculating tools.

The statistics pertaining to fluctuation strength and roughness, on the other hand, are clearly problematic since they are so dissimilar.

4.6 – Comparison of data supplied by Matlab, Testlab, and Brüel for washing machines

In the case of the five washing machines examined, I will only publish statistics from one of them because it is indicative of the entire group.

The comparison was limited to Matlab and Testlab in this example. Furthermore, the articulation index has not been computed because it is not required for the evaluation of a washing machine.

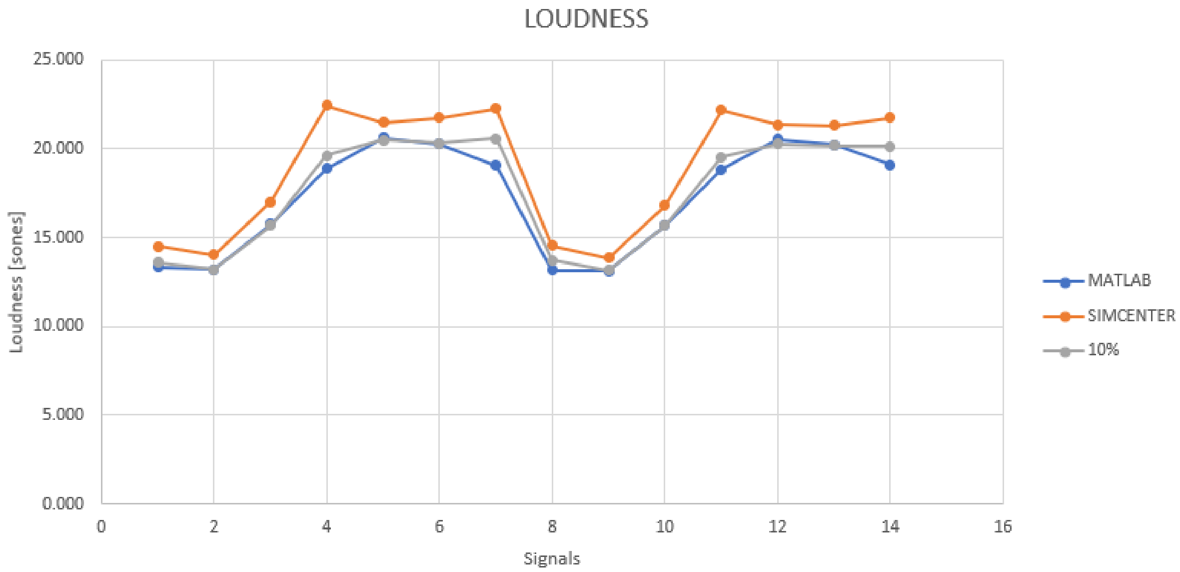

Figure 23: Washing machine loudness comparision between Matlab and Testlab

The trend of the loudness curves produced using Matlab and Testlab (with two alternative parameter values) is similar, as shown in figure 23, and the results, while not entirely consistent, are comparable.

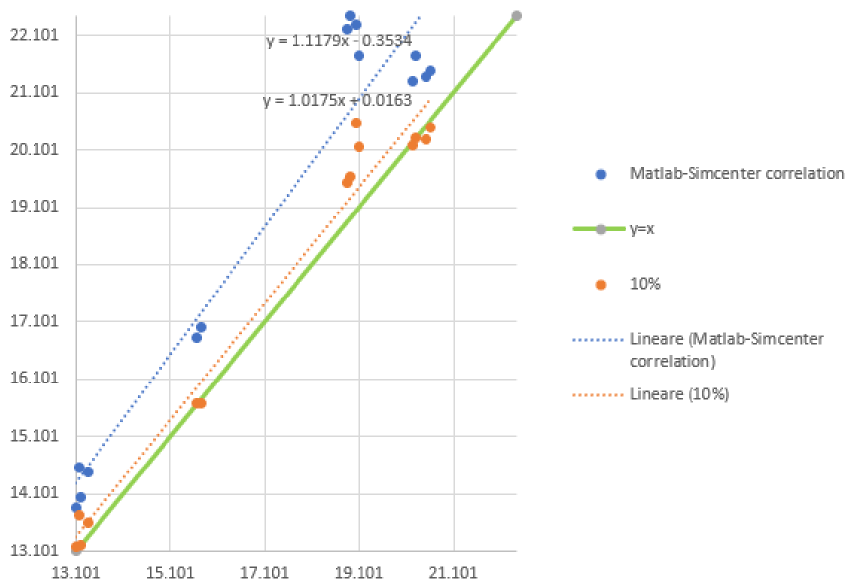

Figure 24: Matlab-Simcenter correlation on loudness data

Figure 24 shows that the data supplied by Matlab and those provided by Testlab have a nearly linear relationship.

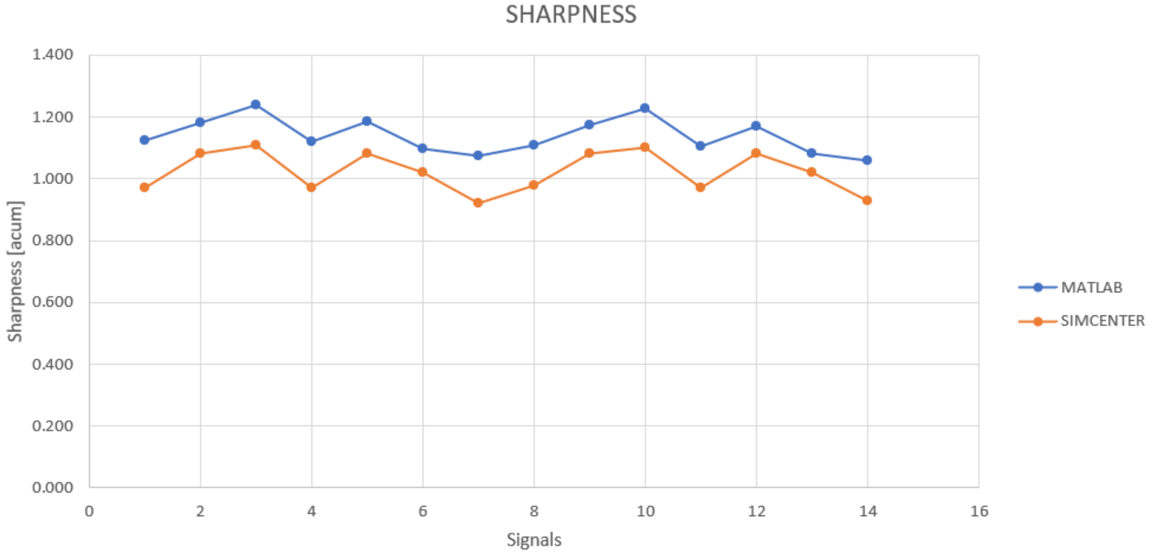

Figure 25: Washing machine sharpness comparision between Matlab and Testlab

Figure 25 shows an unsatisfactory consistency between the sharpness data supplied by the two software programs; nonetheless, they are on the same order of magnitude and follow a similar pattern (the correlation is clear in figure 26).

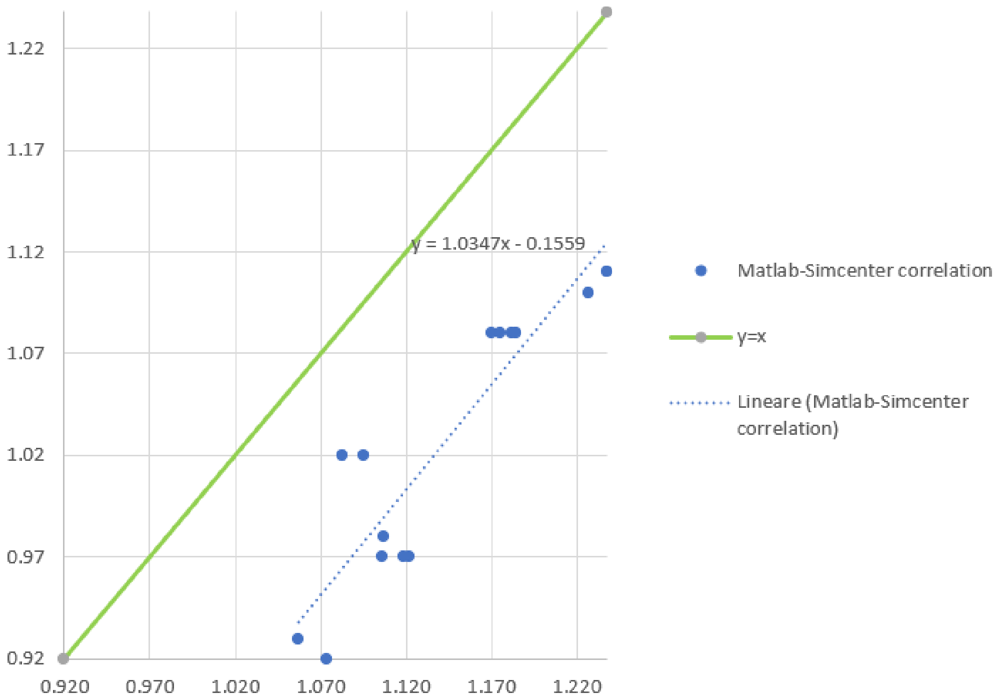

Figure 26: Matlab-Simcenter correlation on sharpness data

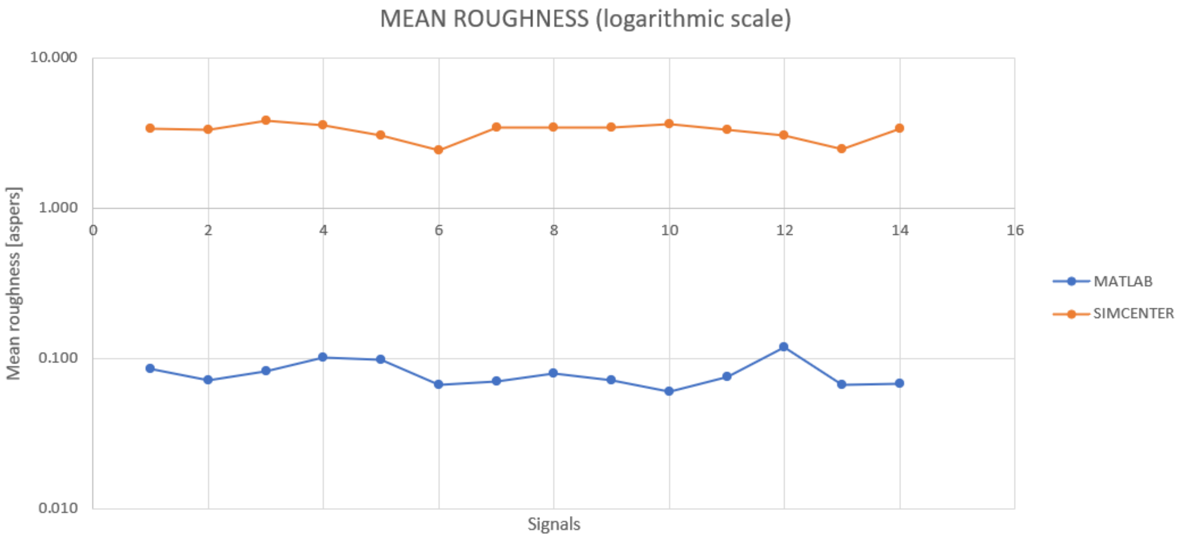

Figure 27: Washing machine roughness comparision between Matlab and Testlab

The large disparity between the two data sets is instantly evident (in figure 27) when it comes to the data pertaining to the roughness calculated using the two techniques.

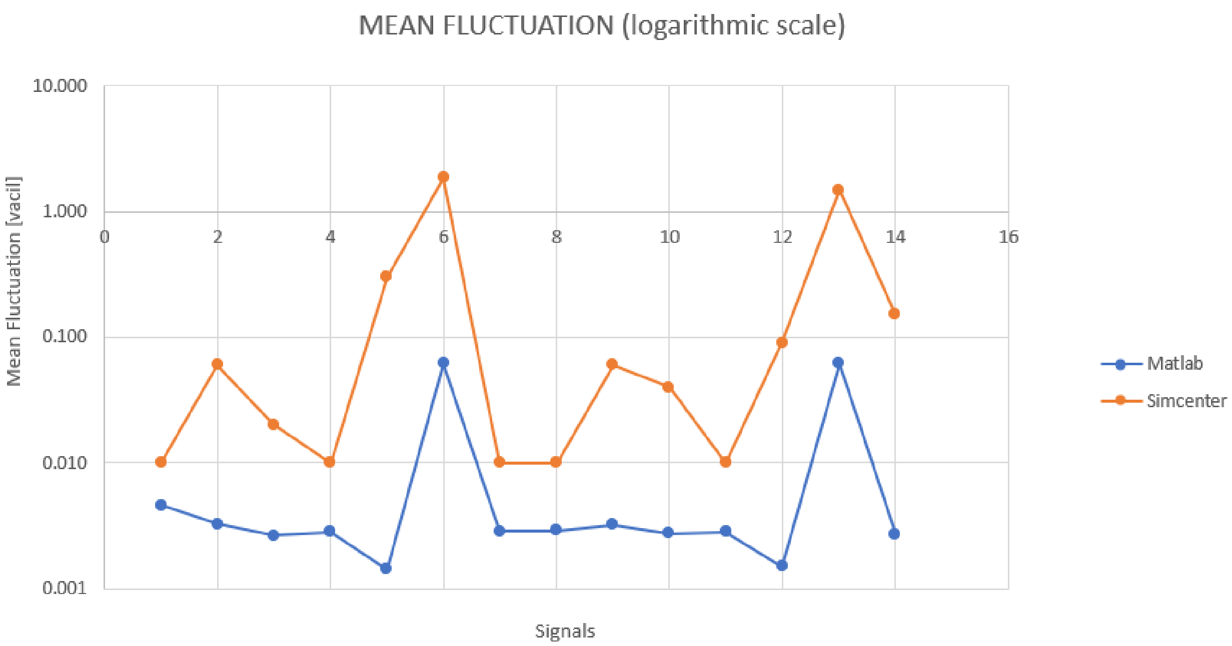

Figure 28: Washing machine fluctuation comparision between Matlab and Testlab

It is feasible to detect a disparity between the two data sets when evaluating the fluctuation strength. However, as seen in figure 28, the two curves appear to have a similar tendency.

5 – Conclusions

Finally, the values of the metrics produced by the various calculating software show obvious variations.

The results obtained for roughness and fluctuation strength, in particular, are very irregular and hence cannot be regarded entirely dependable.

The results collected for the articulation index, loudness, and sharpness are all consistent and hence considered trustworthy.

The major conclusion in terms of the link with the subjective data supplied by the respondents is that as the roughness or fluctuation strength rises, the judgement deteriorates.

The technique outlined in this thesis might be seen as a source of inspiration for future studies using a larger dataset to allow neural networks to be used more effectively.

Bibliography

1. **Siemens,** Sound Quality Metrics: Loudness and Sones

2. **Siemens,** Sharpness in Simcenter Testlab

3. **Siemens,** Sound Modulation Metrics: Fluctuation Strength and Roughness

4. **Siemens,** Articulation Index

5. **Mathworks,** What is a neural network?

6. **Stefano Pagnotta**, Applicazione di reti neurali a casi reali

7. **David L. Bowen**, Correlating Sound Quality Metrics and Jury Ratings